UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

# Projet Économétrie Linéaire

Mélanie Daddio, Ezequiel Hurtado, Ghalia Jaziri

Master 2 Économétrie et Statistiques - TIDE

# Contents

# 1 Presentation du problème et des objectifs de l'étude

Tous les jours, les entreprises sont confrontées au fait de chercher la personne la plus apte pour un poste disponible. Les moyens pour pallier ce besoin sont divers. Ainsi, les équipes cherchent à trouver la meilleure personne en minimisant le coût et le temps d'embauche et en maximisant, bien sûr, la qualité.

Ces enjeux, qui n'en sont pas les seuls, s'inscrivent dans un monde du travail qui évolue rapidement et qui voit naître d'autres questions qui auparavant étaient peu posées . Par exemple, l'inclusion et le fait de s'adapter aux dynamiques des nouvelles générations. À cela, nous pouvons ajouter la pénurie des personnes qualifiées dans certains métiers et le rôle grandissant de l'intelligence artificielle.

En ce qui nous concerne, nous tenterons de répondre au fait d'optimiser ce processus. Cette quête est tout autant intéressante lorsque, comme dit précédemment, les moyens sont nombreux. Nous chercherons ainsi à analyser les données dont nous disposons l'accès. Nous tiendrons à optimiser le processus d'embauche du point de vue de trois points énoncés. Coût, temps et qualité. Ce, dans un cadre analytique et prédictif.

# 2 Presentation des données

Dans cette première période, nous pouvons dire que notre "dataset" contient **11 variables** et **64 observations**. En ce qui concerne les valeurs manquantes, nous n'en avons pas.

En termes de types de données, nous pouvons les diviser en quantitatives et qualitatives de la manière suivante:

- Variables Quantitatives
  - Sl.No : Index de la base de données énumérant les observations.
    - Moyenne: 32.5 et Écart-type: 18.62

– Fiscal Year: Année où l'embauche se déroule. Nous n'avons qu'une seule modalité, 2018.
  * Moyenne : 2018 et Écart-type : 0
– Quarter : Trimestre où la demande d'embauche a été prononcée.
  * Moyenne : 2.70 et Écart-type : 1.06
– Yearly PayScale (Yen) : Salaire de la personne embauchée en année et en Yen.
  * Moyenne : 4.914450e+05 et Écart-type : 2.074369e+05
– **Cost of Hire**[1] : Coût de l'embauche.
  * Moyenne : 18 032 et Écart-type 31 730.56
– **Time of Hire (Days)** : Différence de temps entre la date de parution du besoin et la date où le besoin a été rempli.
  * Moyenne : 35.31 et Écart-type : 13.03
– **Quality of Hire** : Métrique de qualité de l'embauche.
  * Moyenne : 79.44 et Écart-type : 16.27
– Engagement : Métrique d'investissement personnel de la personne embauchée.
  * Moyenne : 50.23 et Écart-type : 21.77
– Ramp up Time : Montée en puissance de compétences au sein de l'entreprise.
  * Moyenne : 76.81 et Écart-type : 22.95
– Culture fit (%) : Métrique montrant l'adaptation en termes culturels au sein de l'entreprise.
  * Moyenne : 82.11 et Écart-type : 12.81

| | Sl. No | Fiscal Year | Quarter | Yearly PayScale (Yen) | Cost_of_Hire | Time_to_Hire_Days | Quality_of_Hire | Engagement | Ramp Up Time | Culture Fit (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 64.000000 | 64.0 | 64.000000 | 6.400000e+01 | 64.000000 | 64.000000 | 64.000000 | 64.000000 | 64.000000 | 64.000000 |
| mean | 32.500000 | 2018.0 | 2.703125 | 4.914450e+05 | 18032.500000 | 35.312500 | 79.437500 | 50.234375 | 76.812500 | 82.109375 |
| std | 18.618987 | 0.0 | 1.064278 | 2.074369e+05 | 31730.556101 | 13.033382 | 16.271384 | 21.767707 | 22.953282 | 12.812590 |
| min | 1.000000 | 2018.0 | 1.000000 | 1.680000e+05 | 2625.000000 | 20.000000 | 41.000000 | 26.000000 | 24.000000 | 47.000000 |
| 25% | 16.750000 | 2018.0 | 2.000000 | 3.405000e+05 | 3656.250000 | 27.750000 | 81.500000 | 34.750000 | 67.250000 | 84.000000 |
| 50% | 32.500000 | 2018.0 | 3.000000 | 4.773000e+05 | 10000.000000 | 31.000000 | 85.000000 | 41.000000 | 84.000000 | 85.000000 |
| 75% | 48.250000 | 2018.0 | 4.000000 | 6.600000e+05 | 10000.000000 | 34.250000 | 89.000000 | 71.000000 | 91.250000 | 87.000000 |
| max | 64.000000 | 2018.0 | 4.000000 | 1.008000e+06 | 120960.000000 | 88.000000 | 94.000000 | 96.000000 | 105.000000 | 95.000000 |

Figure 1: Statistiques descriptives des variables quantitatives

[1]En gras, nous avons les variables cibles.

- Variables Qualitatives
  - Departement : Désigne le département de l'entreprise qui exprime le besoin.
  - Job Open : Date où la demande d'embauche a été prononcé.
  - Hire Date : Date d'embauche.
  - Job Title : Intitulé du poste.
  - Source of Hire : Canal par lequel la demande a été remplie. LinkedIn, Reférences, etc.
  - Performace Score : Variable qualitative oridnale montrant le dégré de performance de la personne embauchée.
  - Sex : Genre de l'individu.

## 2.1 Analyse Univarié

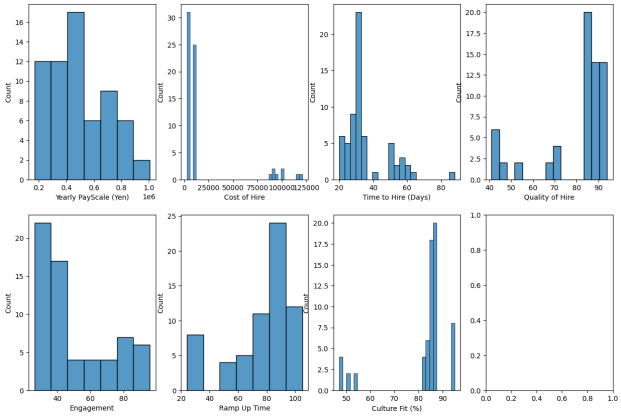



Figure 2: Analyse univarié des variables quantitatives

Lorsque nous étudions les variables quantitatives, nous pouvons nous rendre compte que les salaires se concentrent vers la gauche. Ceci montre que les salaires hauts sont attribués. Ceci est de même pour les coûts d'embauche.

Nous voyons que l'engagement semble bas et que la montée en puissance

semble avoir des valeurs importantes. Le temps d'embauche est aussi assez concentré vers la gauche. Nous voyons aussi finalement que l'adaptation culturelle semble assez grande.

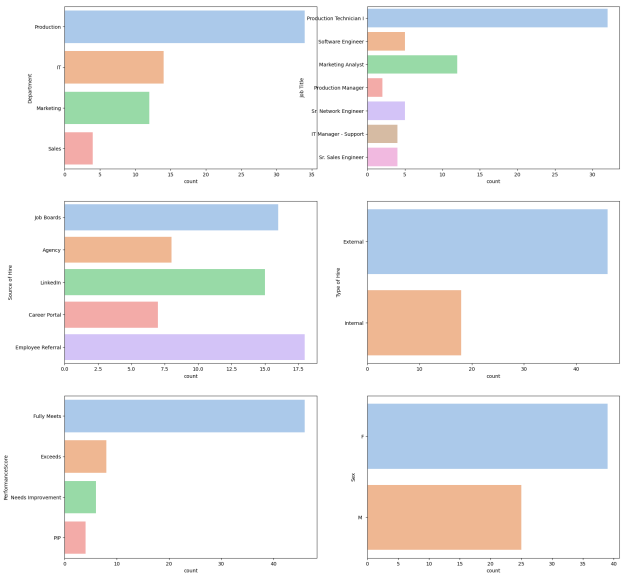



Figure 3: Analyse univarié des variables qualitatives

Les variables quantitatives quant à elles nous montrent, par exemple, que le moyen principal d'embauche semble être le réseautage ou ce qu'ils appellent le "Job Board". Ce sont les équipes de production et IT qui ont embauché le plus et la plupart des embauchés est de genre féminin. Finalement, la plupart des embauchés viennent de l'exterieur de l'entreprise.

La majorité des personnes étaient des femmes ou encore que la source de ces employés était dans la plupart les réferences personlles ou LinkedIn. Cependant dans la plupart, les employés proviennent de l'exterieur de l'entreprise. Le departement ayant eu le plus d'embauches est le département de production suivi par le departement IT.

## 2.2 Analyse Bivarié

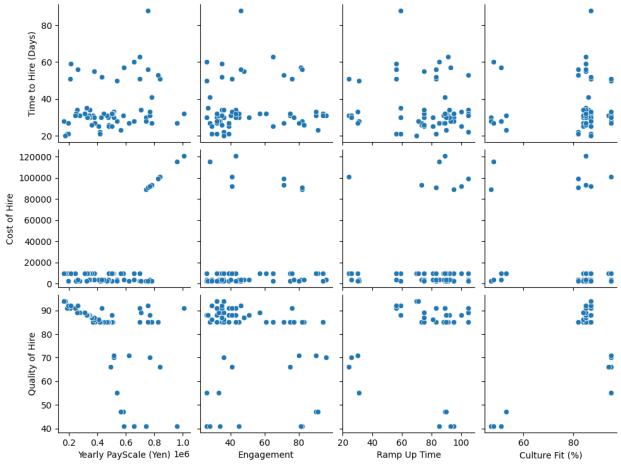



Figure 4: Analyse Bivarié Quantitatif

Dans cette analyse bivarié nous reaffirme ce que nous avons vu avant. Nous avons une densité plus forte vers les haut des graphiques pour la qualité, vers le bas pour les coûts et un comportement plus hétérogène pour le temps d'embauche.

Sur Time to Hire , nous voyons que plus le temps d'embauche est court, plus l'engagement semble bas. Cela se reproduit aussi avec la variable de Salaire Annuel. Avec Cost of Hire on voit une sorte de relation entre Salaire et Coût d'embauche au-délà d'un seuil et cela peut en quelque sorte être l'élasticité.

Lorsque la montée en puissance prend du temps , nous avons une relation avec la qualité d'embauche. L'engagement a un comportement anormale. Cependant, il faut noter que nous faisons que des relations linéaires et elles peuvent vraisemblablement ne pas l'être.

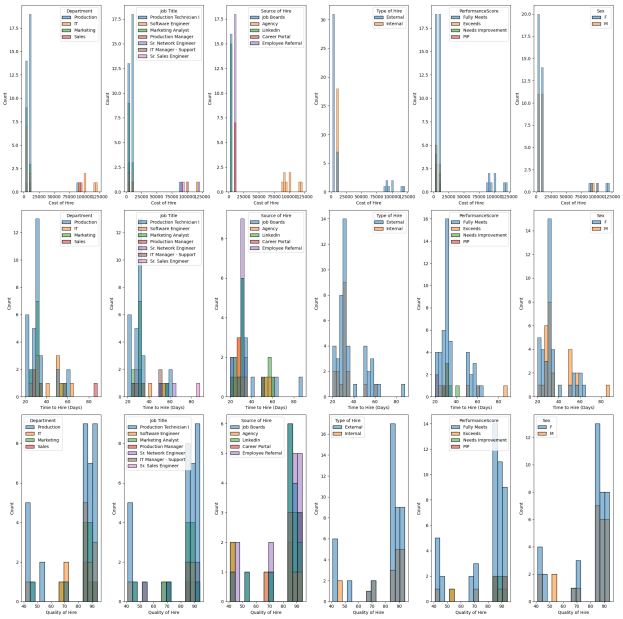


Figure 5: Analyse Bivarié Quantitatif

Bien que peu lisible du fait que les couleurs se superposent, ici, nous tentons de voir l'intéraction entre nos variables cibles et les différentes variables catégorielles.

Par exemple, nous voyons que les coûts d'embauche sont élevés pour l'IT. Pourtant, le titre ayant le plus de coût correspond au Senior Sales Engineer. Les coûts semblent élevés pour ceux qui viennent des agences.

En ce qui concerne le temps d'embauche, nous voyons que l'embauche ayant pris le plus de temps répond à un besoin de l'équipe marketing. Le fait que les couleurs se chevauchent nous montre que les comportements ne sont pas forcément différenciables à cette première étape. Cette analyse est de même pour la variable de qualité de l'embauche.

## 2.3 Analyse Multivarié

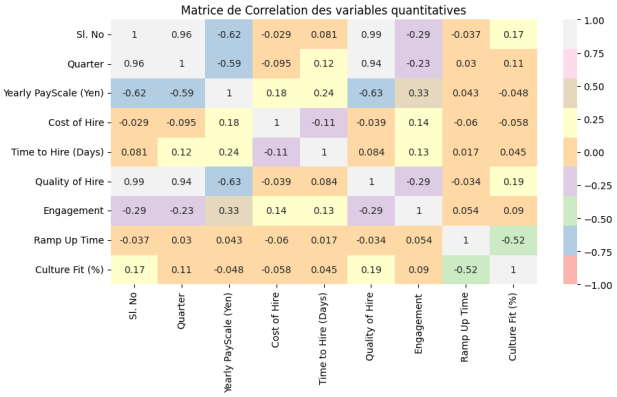



Figure 6: Analyse Multivarié

Dans cette matrice de corrélation, nous apercevons que les variables peu "pertinentes" sont encore là car le nettoyage ou traitement n'a pas encore été fait. Nous nous intéresserons surtout à ce qui semblerait être significatif vis-à-vis à nos variables cibles.

Nous voyons dans un premier abord que le coût d'embauche est corrélé positivement au salaire annuel et négativement avec le temps d'embauche. Si nous continuons, la qualité d'embauche est fortement corrélé avec le Nombre de salarié et Trimestre ce qui est rare. Elle semble être corrélée négativement au salaire annuel et l'engagement et positivement avec l'adaptation culturelle.

Finalement, le temps d'embauche est corrélé avec le salaire annuel et négativement avec le coût d'embauche. Il est important de remarquer que les relations sont faibles et cela laisse penser qu'il existe une possibilité que la relation ne soit pas linéaire.

En étudiant la corrélation entre variables qualitatives, nous pouvons remarquer que la plupart des variables semblent être indépendantes. Tout de même, nous pouvons montrer des relations qui semblent évidentes. Par exemple, entre "Département" et "Job Title".

# 3 Préparation des données

Afin de créer notre modèle, nous sommes passées par une étape de préparation des données. Dans un premier temps, nous avons étudié la normalité des variables. La plupart des variables s'adaptent bien à la loi normale sauf le coût d'embauche et l'adaptation culturelle avec un **W de Shapiro** de 0.49 et 0.62 respectivement.

Par la suite, nous avons décidé de standardiser des données car cela garde les distributions normales et respecte aussi les outliers. Cette méthode est robuste pour les analyses statistiques.

Afin de réaliser cette méthode, nous avons utilisé la méthode de Standard Scaler. Avant de continuer, nous avons réalisés une deuxième matrice de corrélation.

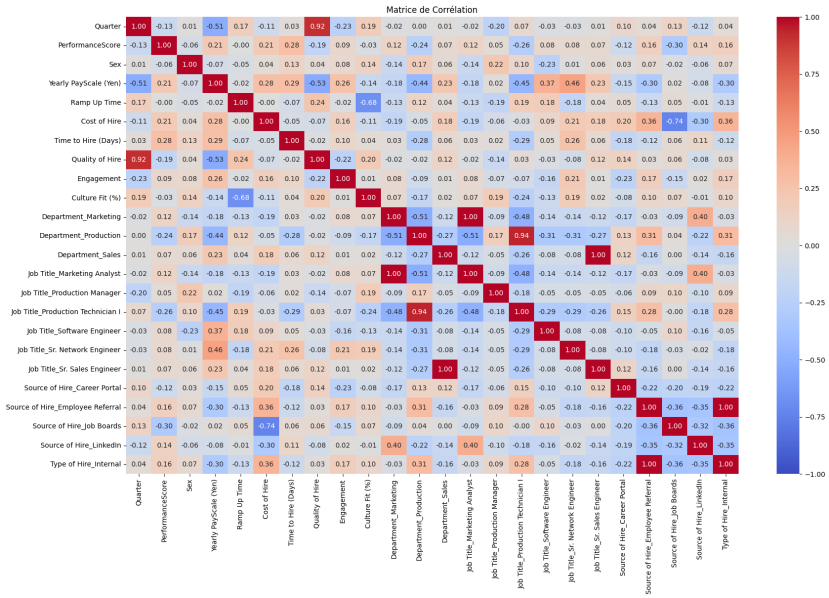



Figure 7: Matrice de corrélation après régularisation

**Cost of Hire** a une corrélation positive avec Yearly PayScale (Yen). Elle est à la hauteur de 0,28. De même manière, elle est corrélée avec le fait que l'employé soit référé et que ce soit un recrutement en interne. En termes de corrélation négative, la plus forte que nous trouvons est de -0,74 avec la

variable "Source of Hire - Job Boards".

Quant à **Quality of Hire**, elle est très corrélée à la variable "Quarter" (0,92). En termes de corrélations positives, elle est suivie par "Ramp Up Time". La corrélation négative la plus forte est à la hauteur de -0,53 avec Yearly PayScale (Yen).

**Time to Hire (Days** n'a pas de corrélation forte. Les grandeurs sont entre +0,29 et -0,29. Du côté des corrélations positives, nous avons Yearly PayScale (Yen) et PerformanceScore. De l'autre côté, nous avons le département de production et le poste de Technicien I.

Lorsque nous analysons l'ensemble de variables explicatives, nous pouvons voir que "Ramp Up Time" est corrélé négativement avec Culture Fit. Cela est normal car plus une personne est adaptée culturellement, plus la personne aura des facilités à monter en compétences.

Pour que la variable ait un impact entre la variable cible et la variable explicative, il faut qu'il y ait une corrélation entre 0,3 et 0,7. Que ce soit positive ou négative. Nous tenons compte aussi qu'une corrélation supérieure peut être un indice de colinéarité.

Nous avons aussi réalisé le calcul de la VIF. Nous avons des scores importants pour des variables telles que "Source of Hire LinkedIn", "Cost of Hire" ou "source of Hire Job Boards"

Afin de standardiser nos variables, nous avons testé plusieurs méthodes. Notamment, la méthode de Box-Cox, Log, Sqrt et la standardisation basique. Cependant, la méthode gardée est la classique donnée par Scikit-Learn car c'était celle qui performait le mieux au vu des tests réalisés. Cette standardisation nous permet interpreter les différentes variables et ne pas avoir des grands écarts en termes d'ordre de grandeur.

De plus, en ce qui concerne les variables qualitatives, lorsque les modalités sont à deux, nous les avons encodées de façon binaire. Pour les variables ayant plus de deux modalités, comme par exemple "Job Title", nous avons utilisé les méthodes de One Hot Encoding. Ceci consiste à faire que chaque modalité devienne une variable sauf une pour ne pas créer de la multicolinéarité.

# 4 Analyse de la Variance (ANOVA)

Dans le cadre de cette étude, nous avons réalisé des tests ANOVA afin d'évaluer l'influence des variables explicatives sur trois variables cibles : **Quality of Hire**, **Cost of Hire** et **Time to Hire**.

L'objectif de ces tests est de déterminer si les différences observées entre les groupes définis par les variables explicatives sont statistiquement significatives. L'ANOVA permet de comparer la variance expliquée par le modèle à la variance résiduelle et d'évaluer si l'effet des variables indépendantes est significatif sur chacune des variables dépendantes.

## 4.1 Méthodologie

Nous avons réalisé deux types de tests ANOVA :

- **ANOVA à un facteur** : Ce test permet d'évaluer l'effet d'une seule variable explicative sur chaque variable cible. Il vise à déterminer si les moyennes des groupes définis par ce facteur sont significativement différentes.
- **ANOVA à deux facteurs** : Cette approche nous permet d'étudier l'influence conjointe de deux variables explicatives sur la variable cible. En plus d'analyser l'effet individuel de chaque facteur, nous évaluons également leur interaction.

Nous avons utilisé un modèle de régression linéaire et appliqué un test ANOVA pour examiner l'impact global des variables explicatives sur chaque variable cible. Le test repose sur les hypothèses suivantes :

- $H_0$ **(hypothèse nulle)** : Les moyennes des groupes sont identiques, indiquant que la variable explicative n'a pas d'effet significatif sur la variable cible.
- $H_1$ **(hypothèse alternative)** : Au moins un groupe présente une différence significative, suggérant une influence des variables explicatives.

## 4.2 Résultats des Tests ANOVA pour la Variable *Time_To_Hire_Days*

Les résultats des tests ANOVA montrent que plusieurs variables explicatives ont un effet significatif sur la variable cible *Time_to_Hire_Days*. Les variables suivantes présentent des différences significatives :

- **PerformancesScore** (p-value : 0.022556) : Cette variable a un effet significatif sur le temps de recrutement, ce qui suggère que les performances des candidats influencent la durée du processus de recrutement.
- **Department_Production** (p-value : 0.025325) : Le département de production a un impact significatif sur le temps de recrutement, indiquant que les processus de recrutement dans ce département sont différents des autres départements.
- **Job_Title_Production_Technician_I** (p-value : 0.021247) : Une différence significative a été observée pour ce titre de poste par rapport aux autres. Cela suggère que le processus de recrutement pour ce rôle est distinct.
- **Job_Title_Sr_Network_Engineer** (p-value : 0.035525) : Il existe une différence significative pour ce titre de poste par rapport aux autres, ce qui indique que les pratiques de recrutement varient pour ce rôle spécifique.

## 4.3 Interactions de Variables pour la Variable *Time_To_Hire_Days*

Les résultats des tests ANOVA pour les interactions entre les variables montrent que plusieurs interactions ont un effet significatif sur la variable cible *Time_to_Hire_Days*. Les interactions suivantes présentent des différences significatives :

- **C(Department_Sales):C(Source_of_Hire_Job_Boards)** (p-value : 0.000044) : Cette interaction montre une influence significative entre le département des ventes et la source de recrutement via les site d'emplois sur le temps de recrutement.
- **C(Department_Sales):C(PerformanceScore)** (p-value : 0.000318) : L'interaction entre le département des ventes et le score de performance a un effet significatif, suggérant que les performances des candidats influencent différemment le temps de recrutement dans ce département.

- **(Job_Title_Sr_Sales_Engineer):C(Source_of_Hire_Job_Boards)** (p-value : 0.000044) : Il existe une interaction significative entre le titre de poste de *Sr Sales Engineer* et la source de recrutement via les site d'emplois, influençant ainsi le temps de recrutement.

- **C(Job_Title_Sr_Sales_Engineer):C(PerformanceScore)** (p-value : 0.000318) : L'interaction entre le titre de poste de *Sr Sales Engineer* et le score de performance montre un effet significatif, indiquant que le score de performance a un impact particulier sur le processus de recrutement pour ce poste.

## 4.4 Résultats des Tests ANOVA pour la Variable *Cost_of_Hire*

Les tests ANOVA indiquent que les variables suivantes ont un effet significatif sur *Cost_of_Hire* :

- **Source_of_Hire_Employee_Referral** (p-value: **0.00389**)
- **Source_of_Hire_Job_Boards** (p-value: **4.46e-12**)
- **Source_of_Hire_LinkedIn** (p-value: **0.016016**)
- **Type_of_Hire_Internal** (p-value: **0.00389**)

Ces résultats suggèrent que la source de recrutement ainsi que le type d'embauche influencent significativement le coût d'embauche.

## 4.5 Résultats des Tests ANOVA pour la Variable *Quality_of_Hire*

Aucune des variables testées n'est statistiquement significative au seuil de 5 % ($p > 0.05$).

**Test le plus proche de la significativité :**

- **PerformanceScore** ($F = 2.21$, $p = 0.14175$) : Bien que cette variable soit la plus proche du seuil de significativité, elle reste non significative et ne permet pas de conclure à un effet sur *Quality_of_Hire*.

## 4.6 Interactions de Variables pour la Variable *Quality_of_Hire*

Les résultats des tests ANOVA montrent que plusieurs interactions ont un effet significatif sur la variable cible *Quality_of_Hire*. Les interactions suivantes sont statistiquement significatives ($p < 0.05$) :

- **C(Department_Marketing):C(Source_of_Hire_Employee_Referral)** (p-value : 0.001289) : L'interaction entre le département marketing et la source de recrutement via la recommandation interne a un effet significatif sur la qualité des recrutements.
- **C(Department_Marketing):C(Type_of_Hire_Internal)** (p-value : 0.001289) : Cette interaction indique que le recrutement interne impacte significativement la qualité des recrutements au sein du département marketing.
- **C(Department_Production):C(Type_of_Hire_Internal)** (p-value : 0.007534) : L'effet du recrutement interne sur la qualité des recrutements est également significatif pour le département de production.
- **C(Department_Production):C(Source_of_Hire_Employee_Referral)** (p-value : 0.007534) : La recommandation interne a un effet significatif sur la qualité des recrutements dans le département de production.

# 5 Impact des variables cibles

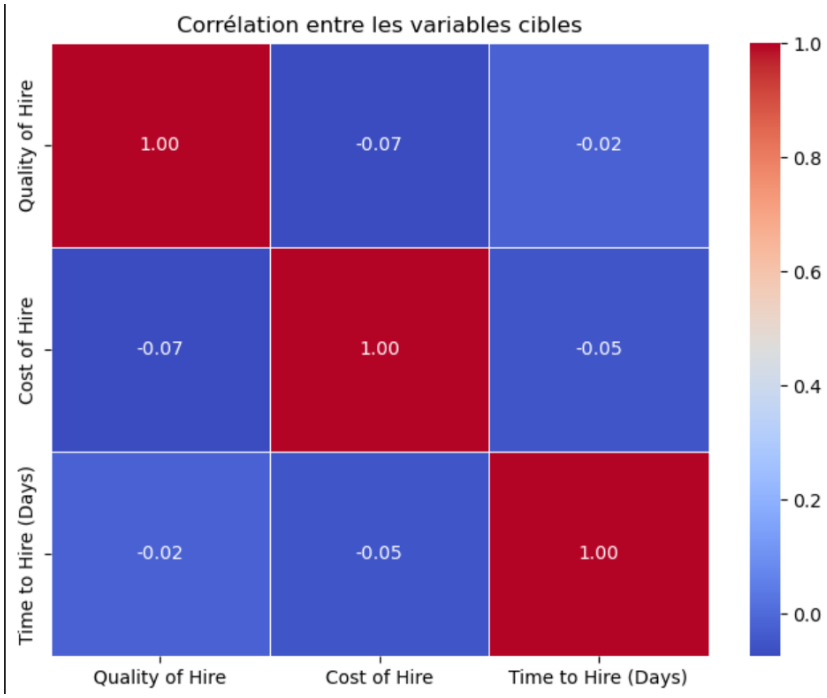



Figure 8: Matrice de Corrélation des variables cibles

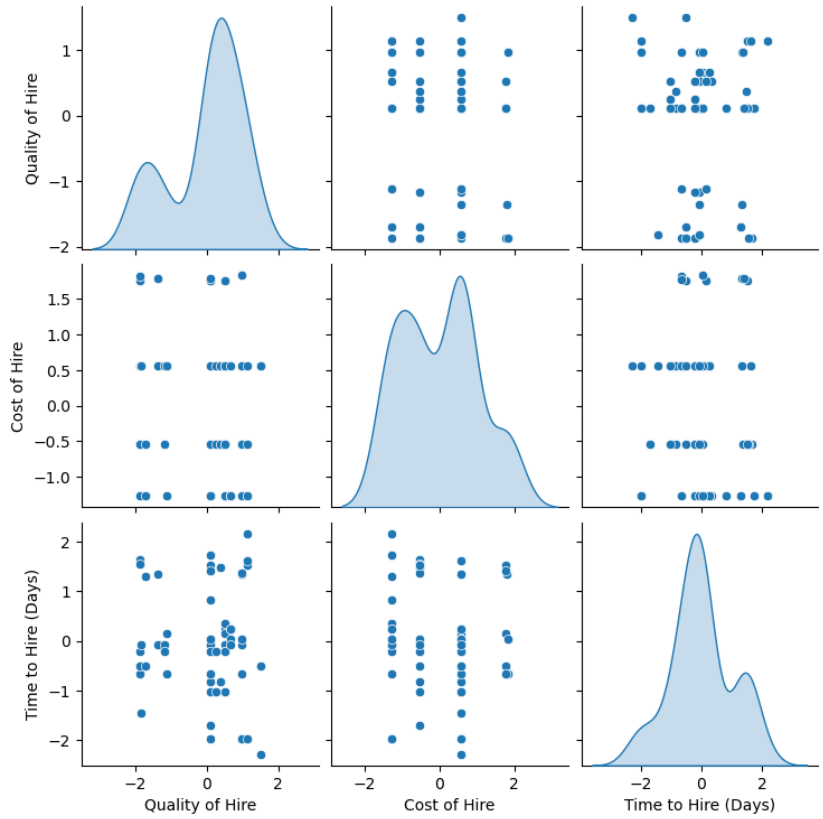



Figure 9: Lien entre les variables cibles

L'analyse des corrélations entre les variables **Quality of Hire**, **Cost of Hire** et **Time to Hire** met en évidence les relations suivantes :

- **Quality of Hire & Cost of Hire** : La corrélation est de **-0.07**, indiquant une très légère relation négative. Cependant, cette relation est négligeable et ne permet pas de conclure à un lien significatif entre la qualité d'embauche et son coût.
- **Quality of Hire & Time to Hire** : La corrélation est de **-0.02**, ce qui signifie qu'il n'existe pratiquement aucune relation entre la qualité d'embauche et le temps de recrutement.
- **Cost of Hire & Time to Hire** : La corrélation est de **-0.05**, indiquant une très faible relation négative, qui n'est pas significative d'un point de vue statistique.

Ces résultats suggèrent que **le coût et le temps de recrutement n'ont pas d'impact notable sur la qualité de l'embauche** dans notre modèle.

# 6 Régression Linéaire

La régression linéaire est l'une des techniques fondamentales en statistique et en apprentissage automatique, utilisée pour modéliser la relation entre une variable dépendante continue et une ou plusieurs variables indépendantes. Elle repose sur l'idée que cette relation peut être représentée par une équation linéaire, où la variable cible est une fonction linéaire des variables explicatives.

Le but principal de la régression linéaire est d'estimer les paramètres d'une droite (ou hyperplan dans le cas de variables multiples) qui minimise l'écart entre les valeurs observées et celles prédites par le modèle. Bien que simple et facile à interpréter, la régression linéaire forme la base pour des techniques plus avancées qui permettent de traiter des relations plus complexes et de prendre en compte davantage de variables.

Dans cette section, nous explorerons les principes de base de la régression linéaire avant de passer à des modèles plus sophistiqués qui étendent cette approche pour capturer des relations plus complexes et non linéaires entre les variables.

Nous avons commencer par faire un modèle complet pour les trois variables cibles et nous avons pris toute les variables donnée dans la Data Frame excepté les variables de temps (Fiscal Year, Job Open Date et Hire Date).

## 6.1 *Cost_of_Hire*

Nous avons fait un modèle complet pour la variable Cost_of_Hire.

P-values pour la régression de Cost_of_Hire :

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Cost_of_Hire | R-squared: | 1.000 |
| Model: | OLS | Adj. R-squared: | 1.000 |
| Method: | Least Squares | F-statistic: | 4.081e+04 |
| Date: | Wed, 05 Mar 2025 | Prob (F-statistic): | 3.18e-90 |
| Time: | 04:14:59 | Log-Likelihood: | 217.07 |
| No. Observations: | 64 | AIC: | -398.1 |
| Df Residuals: | 46 | BIC: | -359.3 |
| Df Model: | 17 | | |
| Covariance Type: | nonrobust | | |

Figure 10: Résultat du modèle 1 complet pour `Cost_of_Hire`

Le modèle présente un $R^2$ de 1.000, indiquant que le modèle explique parfaitement la variance de la variable dépendante. Le $R^2$ ajusté est également de 1.000, ce qui confirme que le modèle s'ajuste bien aux données sans surajustement. La statistique $F$ est de 4.081e+04 avec une probabilité associée de $3.18 \times 10^{-90}$, ce qui indique que le modèle est hautement significatif dans son ensemble.

Le nombre d'observations est de 64, et le nombre de degrés de liberté des résidus est de 46. Les critères d'information d'Akaike (AIC) et de Bayes (BIC) sont respectivement de -398.1 et -359.3, ce qui suggère que le modèle est bien adapté par rapport à d'autres modèles potentiels.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 1.7820 | 0.009 | 193.939 | 0.000 | 1.763 | 1.800 |
| Culture_Fit_ | -0.0002 | 0.002 | -0.109 | 0.914 | -0.004 | 0.004 |
| Department_Marketing | 0.0020 | 0.003 | 0.660 | 0.513 | -0.004 | 0.008 |
| Department_Production | 0.0020 | 0.005 | 0.445 | 0.658 | -0.007 | 0.011 |
| Department_Sales | -0.0045 | 0.004 | -1.223 | 0.227 | -0.012 | 0.003 |
| Engagement | -0.0033 | 0.001 | -2.277 | 0.027 | -0.006 | -0.000 |
| Job_Title_Marketing_Analyst | 0.0020 | 0.003 | 0.660 | 0.513 | -0.004 | 0.008 |
| Job_Title_Production_Manager | 0.0018 | 0.005 | 0.327 | 0.745 | -0.009 | 0.013 |
| Job_Title_Production_Technician_I | 0.0003 | 0.003 | 0.073 | 0.942 | -0.007 | 0.007 |
| Job_Title_Software_Engineer | 0.0023 | 0.008 | 0.303 | 0.764 | -0.013 | 0.017 |
| Job_Title_Sr_Network_Engineer | 0.0046 | 0.007 | 0.622 | 0.537 | -0.010 | 0.020 |
| Job_Title_Sr_Sales_Engineer | -0.0045 | 0.004 | -1.223 | 0.227 | -0.012 | 0.003 |
| PerformanceScore | -0.0040 | 0.006 | -0.709 | 0.482 | -0.016 | 0.007 |
| Quarter | 0.0030 | 0.002 | 1.582 | 0.120 | -0.001 | 0.007 |
| Ramp_Up_Time | -0.0001 | 0.002 | -0.077 | 0.939 | -0.004 | 0.004 |
| Sex | -0.0010 | 0.003 | -0.365 | 0.717 | -0.006 | 0.004 |
| Source_of_Hire_Career_Portal | -1.2095 | 0.006 | -189.249 | 0.000 | -1.222 | -1.197 |
| Source_of_Hire_Employee_Referral | -0.6036 | 0.003 | -196.852 | 0.000 | -0.610 | -0.597 |
| Source_of_Hire_Job_Boards | -3.0558 | 0.005 | -578.045 | 0.000 | -3.066 | -3.045 |
| Source_of_Hire_LinkedIn | -2.3211 | 0.006 | -391.252 | 0.000 | -2.333 | -2.309 |
| Type_of_Hire_Internal | -0.6036 | 0.003 | -196.852 | 0.000 | -0.610 | -0.597 |
| Yearly_PayScale_Yen | 0.0064 | 0.003 | 2.277 | 0.028 | 0.001 | 0.012 |

Figure 11: Résultat du modèle 1 complet pour `Cost_of_Hire` des variables

Dans cette section, nous présentons les résultats significatifs obtenus lors de la régression linéaire pour la variable dépendante `Cost_of_Hire`. Les résultats ont été obtenus en utilisant la méthode des moindres carrés ordinaires (OLS). Seules les variables ayant une valeur de p inférieure à 0.05 ont été considérées comme significatives. Voici les variables et leurs interprétations :

- **Engagement** :

- Coefficient = -0.0033, p-value = 0.027
- La variable `Engagement` a un effet négatif sur le `Cost_of_Hire`. Pour chaque augmentation de 1 unité de l'`Engagement`, le coût d'embauche diminue en moyenne de 0.0033. Cette relation est statistiquement significative au niveau de 5%.

- **Source_of_Hire_Career_Portal** :
  - Coefficient = -1.2095, p-value = 0.000
  - Interprétation : Le recours à un portail de carrière pour le recrutement réduit significativement le `Cost_of_Hire` de 1.2095 en moyenne. Cette relation est hautement significative avec une p-value inférieure à 0.01.

- **Source_of_Hire_Employee_Referral** :
  - Coefficient = -0.6036, p-value = 0.000
  - Interprétation : Le recours à des recommandations d'employés pour le recrutement diminue le coût d'embauche de 0.6036 en moyenne. Ce résultat est également hautement significatif.

- **Source_of_Hire_Job_Boards** :
  - Coefficient = -3.0558, p-value = 0.000
  - Interprétation : L'utilisation des sites de recrutement pour le recrutement entraîne une réduction significative du coût d'embauche de 3.0558. Cette variable est fortement significative avec une p-value très faible.

- **Source_of_Hire_LinkedIn** :
  - Coefficient = -2.3211, p-value = 0.000
  - Interprétation : Le recrutement via LinkedIn réduit également le coût d'embauche de 2.3211 en moyenne. Cette variable est aussi hautement significative, montrant l'efficacité de ce canal de recrutement pour réduire les coûts d'embauche.

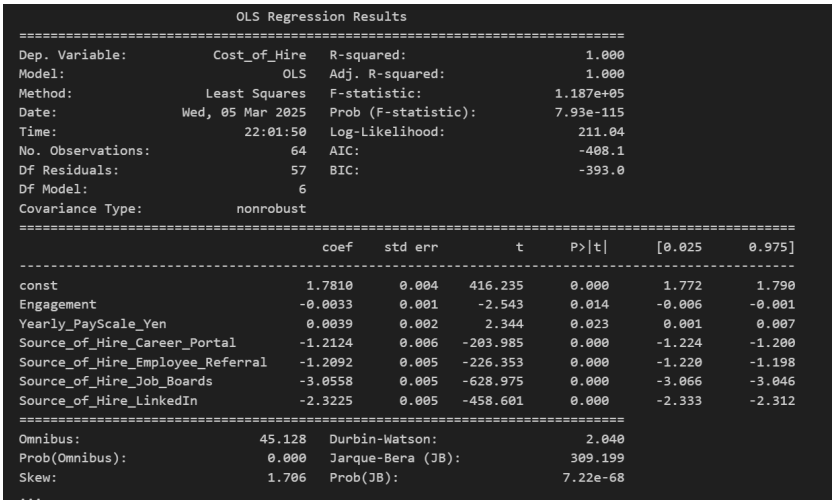
- **Type_of_Hire_Internal** :

- Coefficient = -0.6036, p-value = 0.000
- Interprétation : Le recrutement interne réduit de manière significative le coût d'embauche de 0.6036. Cette relation est hautement significative, suggérant que les recrutements internes sont moins coûteux que les recrutements externes.

- **Yearly_PayScale_Yen** :
  - Coefficient = 0.0064, p-value = 0.028
  - Interprétation : Le salaire annuel en yens (`Yearly_PayScale_Yen`) a un effet positif sur le coût d'embauche, avec un coefficient de 0.0064. Cela signifie qu'à mesure que le salaire annuel augmente, le coût d'embauche augmente également. Cette relation est statistiquement significative au niveau de 5%.

Ces résultats mettent en évidence l'importance des canaux de recrutement et du type de recrutement (interne ou externe) dans la détermination du coût d'embauche. De plus, le salaire annuel a un effet positif sur ce coût, ce qui est logique dans la mesure où un salaire plus élevé implique souvent des coûts d'embauche plus importants.

Nous avons decidé de garder seulement les variables significatifs excepté Type of Hire Internal car elle est linéairement dependante avec Source_of_Hire_Employee_Referral. Nous avons fait un nouveau modèle avec ces mêmes variables.



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Cost_of_Hire | R-squared: | 1.000 |
| Model: | OLS | Adj. R-squared: | 1.000 |
| Method: | Least Squares | F-statistic: | 1.187e+05 |
| Date: | Wed, 05 Mar 2025 | Prob (F-statistic): | 7.93e-115 |
| Time: | 22:01:50 | Log-Likelihood: | 211.04 |
| No. Observations: | 64 | AIC: | -408.1 |
| Df Residuals: | 57 | BIC: | -393.0 |
| Df Model: | 6 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 1.7810 | 0.004 | 416.235 | 0.000 | 1.772 | 1.790 |
| Engagement | -0.0033 | 0.001 | -2.543 | 0.014 | -0.006 | -0.001 |
| Yearly_PayScale_Yen | 0.0039 | 0.002 | 2.344 | 0.023 | 0.001 | 0.007 |
| Source_of_Hire_Career_Portal | -1.2124 | 0.006 | -203.985 | 0.000 | -1.224 | -1.200 |
| Source_of_Hire_Employee_Referral | -1.2092 | 0.005 | -226.353 | 0.000 | -1.220 | -1.198 |
| Source_of_Hire_Job_Boards | -3.0558 | 0.005 | -628.975 | 0.000 | -3.066 | -3.046 |
| Source_of_Hire_LinkedIn | -2.3225 | 0.005 | -458.601 | 0.000 | -2.333 | -2.312 |

| | | | |
|---|---|---|---|
| Omnibus: | 45.128 | Durbin-Watson: | 2.040 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 309.199 |
| Skew: | 1.706 | Prob(JB): | 7.22e-68 |

Figure 12: Résultats du modèle 2 des variables pour `Cost_of_Hire`

Le modèle présente un $R^2$ de 1.000, indiquant que le modèle explique parfaitement la variance de la variable dépendante. Le $R^2$ ajusté est également de 1.000, ce qui confirme que le modèle s'ajuste bien aux données sans surajustement. La statistique F est de 1.187e+05 avec une probabilité associée de $7.93 \times 10^{11}$, ce qui indique que le modèle est hautement significatif dans son ensemble.

Le nombre d'observations est de 64, et le nombre de degrés de liberté des résidus est de 57. Les critères d'information d'Akaike (AIC) et de Bayes (BIC) sont respectivement de -408.1 et -393.0, ce qui suggère que le modèle est bien adapté par rapport à d'autres modèles potentiels.

Toutes les variables affichent des p-valeurs très faibles (proches de 0), ce qui signifie qu'elles sont statistiquement significatives au seuil de 5%. Cela indique que chacune d'elles a un effet important sur le coût d'embauche.

- **Engagement** : Coefficient négatif (-0.0033), ce qui signifie qu'une augmentation de l'engagement des employés est associée à une réduction du coût d'embauche.
- **Yearly_PayScale_Yen** : Coefficient positif (0.0039), indiquant qu'une

augmentation de la rémunération annuelle en yen augmente le coût d'embauche.

- **Source_of_Hire_Career_Portal** : Coefficient négatif (-1.2171), suggérant que les embauches via un Career Portal coûtent moins cher que la catégorie de référence.
- **Source_of_Hire_Employee_Referral** : Coefficient négatif (-0.6085), indiquant que les embauches par références internes réduisent le coût d'embauche.
- **Source_of_Hire_Job_Boards** : Coefficient négatif (-3.0600), ce qui montre que l'embauche via des Job Boards est la moins coûteuse par rapport à la catégorie de référence.
- **Source_of_Hire_LinkedIn** : Coefficient négatif (-2.3284), indiquant que les embauches via LinkedIn coûtent également moins cher.

  Le modèle met en évidence un impact significatif des sources d'embauche, de l'engagement et de la rémunération sur le coût d'embauche. Toutefois, un $R^2$ de 1.000 peut indiquer un surajustement, ce qui nécessite une analyse plus approfondie pour confirmer la robustesse des résultats.

  Pour cela, nous avons decidé de tester differentes hypothéses pour la validité d'un modèle. On a testé l'auto-corrélation des residus, l'homoscédasticité et la normalité des résidu.

```
Statistique de Durbin-Watson : 2.0399 (proche de 2 = pas d'autocorrélation)
Test de Breusch-Pagan - p-value : 0.0001 (p>0.05 = homoscédasticité)

Tests de normalité des résidus :
Shapiro-Wilk p-value : 0.0000 (p>0.05 = normalité)
Kolmogorov-Smirnov p-value : 0.0011 (p>0.05 = normalité)

Vérification de la multicolinéarité :
                          Variable        VIF
0                            const  13.035777
1                       Engagement   1.200579
2                Yearly_PayScale_Yen   1.929583
3      Source_of_Hire_Career_Portal   2.450091
4  Source_of_Hire_Employee_Referral   4.107287
5          Source_of_Hire_Job_Boards   3.151325
6            Source_of_Hire_LinkedIn   3.277001
```

Figure 13: Validité du modèle `Cost_of_Hire`

**1. Autocorrélation des résidus (Durbin-Watson)** : La valeur de Durbin-Watson proche de 2 indique qu'il n'y a pas d'autocorrélation significative des résidus. Cela suggère que les erreurs du modèle sont indépendantes, ce qui est une condition importante pour la validité des résultats de la régression.

**2. Homoscedasticité (Breusch-Pagan)** : La p-value obtenue pour le test de Breusch-Pagan est égale à 0, ce qui indique la présence de hétéroscédasticité dans le modèle. Cela peut être dû à la structure des variables catégorielles et à un éventuel déséquilibre dans les groupes, où certaines modalités ont beaucoup plus d'observations que d'autres. Une telle situation peut affecter l'efficacité des estimateurs des moindres carrés.

**3. Normalité des résidus** : Les tests de normalité des résidus (Shapiro-Wilk et Kolmogorov-Smirnov) ont donné des p-values égales à 0, ce qui entraîne le rejet de l'hypothèse de normalité des résidus. Cependant, avec des variables catégorielles dans le modèle, la normalité des résidus n'est pas toujours une hypothèse essentielle, surtout lorsque l'échantillon est grand, car le théorème central limite permet de relativiser cette contrainte.

**4. Multicolinéarité (VIF)** : Aucun problème de multicolinéarité n'a été détecté, les Variance Inflation Factors (VIF) étant dans une plage acceptable. Cela signifie que les variables indépendantes ne sont pas trop corrélées entre elles, ce qui garantit des estimations fiables des coefficients du modèle.

En conclusion, le deuxième modèle s'avère plus performant, offrant des résultats améliorés avec un meilleur ajustement ($R^2$ de 1.000) et des critères d'Akaike (AIC) et Bayes (BIC) plus faibles (-408.1 et -393.0). De plus, toutes les variables explicatives sont hautement significatives, ce qui renforce la robustesse du modèle. En comparaison, le premier modèle, bien qu'efficace, présente un risque de surajustement. Par conséquent, nous avons choisi de retenir le second modèle pour sa meilleure adéquation aux données et ses variables plus significatives.

## 6.2 *Time_to_Hire_Days*

Nous avons fait un modèle complet pour la variable Time_to_Hire_Days.

P-values pour la régression de Time_to_Hire_Days :

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Time_to_Hire_Days | R-squared: | 0.295 |
| Model: | OLS | Adj. R-squared: | 0.035 |
| Method: | Least Squares | F-statistic: | 1.133 |
| Date: | Wed, 05 Mar 2025 | Prob (F-statistic): | 0.355 |
| Time: | 04:14:59 | Log-Likelihood: | -79.622 |
| No. Observations: | 64 | AIC: | 195.2 |
| Df Residuals: | 46 | BIC: | 234.1 |
| Df Model: | 17 | | |
| Covariance Type: | nonrobust | | |

Figure 14: Résultat du modèle complet pour *Time_to_Hire_Days*

Le modèle présente un $R^2$ de 0.295, indiquant que 29.5% de la variance de la variable dépendante *Time_to_Hire_Days* est expliquée par les variables indépendantes du modèle. Le $R^2$ ajusté est de 0.035, ce qui suggère que l'ajout de certaines variables explicatives pourrait ne pas apporter une réelle amélioration de l'ajustement du modèle aux données.

La statistique $F$ est de 1.133 avec une probabilité associée de 0.355, ce qui indique que, dans son ensemble, le modèle n'est pas statistiquement significatif. Autrement dit, il n'existe pas de preuve suffisante pour affirmer que les variables explicatives sélectionnées influencent significativement *Time_to_Hire_Days*.

Le nombre d'observations est de 64, et le nombre de degrés de liberté des résidus est de 46. Les critères d'information d'Akaike (AIC) et de Bayes (BIC) sont respectivement de 195.2 et 234.1. Ces valeurs relativement élevées suggèrent que d'autres modèles pourraient potentiellement mieux s'adapter aux données.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -1.2715 | 0.947 | -1.342 | 0.186 | -3.179 | 0.636 |
| Culture_Fit_ | -0.2179 | 0.206 | -1.059 | 0.295 | -0.632 | 0.196 |
| Department_Marketing | -0.1161 | 0.315 | -0.369 | 0.714 | -0.750 | 0.518 |
| Department_Production | -0.1419 | 0.464 | -0.306 | 0.761 | -1.076 | 0.792 |
| Department_Sales | -0.0531 | 0.378 | -0.141 | 0.889 | -0.814 | 0.708 |
| Engagement | 0.0333 | 0.148 | 0.224 | 0.824 | -0.265 | 0.332 |
| Job_Title_Marketing_Analyst | -0.1161 | 0.315 | -0.369 | 0.714 | -0.750 | 0.518 |
| Job_Title_Production_Manager | 0.0536 | 0.554 | 0.097 | 0.923 | -1.061 | 1.168 |
| Job_Title_Production_Technician_I | -0.1955 | 0.354 | -0.552 | 0.584 | -0.908 | 0.517 |
| Job_Title_Software_Engineer | -0.2290 | 0.775 | -0.295 | 0.769 | -1.790 | 1.332 |
| Job_Title_Sr_Network_Engineer | 0.2573 | 0.764 | 0.337 | 0.738 | -1.281 | 1.796 |
| Job_Title_Sr_Sales_Engineer | -0.0531 | 0.378 | -0.141 | 0.889 | -0.814 | 0.708 |
| PerformanceScore | 1.0205 | 0.587 | 1.737 | 0.089 | -0.162 | 2.203 |
| Quarter | 0.3157 | 0.194 | 1.626 | 0.111 | -0.075 | 0.706 |
| Ramp_Up_Time | -0.2097 | 0.197 | -1.066 | 0.292 | -0.606 | 0.186 |
| Sex | 0.3880 | 0.276 | 1.405 | 0.167 | -0.168 | 0.944 |
| Source_of_Hire_Career_Portal | 0.0228 | 0.659 | 0.035 | 0.973 | -1.304 | 1.349 |
| Source_of_Hire_Employee_Referral | 0.1492 | 0.316 | 0.472 | 0.639 | -0.487 | 0.786 |
| Source_of_Hire_Job_Boards | 0.5988 | 0.545 | 1.099 | 0.278 | -0.498 | 1.696 |
| Source_of_Hire_LinkedIn | 0.6251 | 0.612 | 1.022 | 0.312 | -0.606 | 1.856 |
| Type_of_Hire_Internal | 0.1492 | 0.316 | 0.472 | 0.639 | -0.487 | 0.786 |
| Yearly_PayScale_Yen | 0.3304 | 0.291 | 1.134 | 0.263 | -0.256 | 0.917 |

Figure 15: Résultat du modèle complet pour *Time_to_Hire_Days*

L'Intercept est négatif ($-1.2715$) mais non significatif ($p = 0.186$).

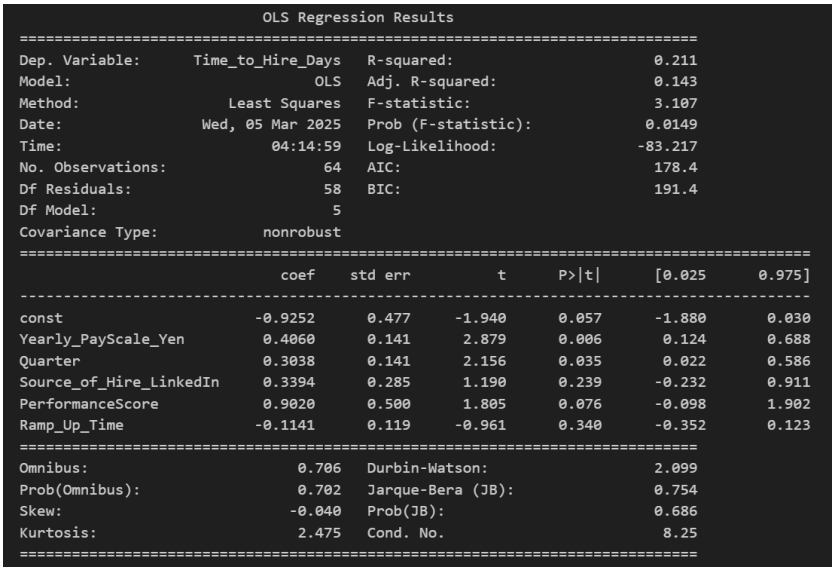
La variable **PerformanceScore** (1.0205, $p = 0.089$) semble avoir l'effet le plus fort, bien qu'elle ne soit significative qu'à environ 10%.
Aucune variable ne semble statistiquement significative au seuil de 5%, ce qui signifie qu'aucune n'a un effet suffisamment fort pour être confirmée avec un haut degré de confiance.

- **Sex** (0.3880, $p = 0.167$) : L'effet du sexe est positif mais non

significatif.

- **Source of Hire** : Aucune des sources d'embauche ne semble avoir un impact significatif.
- **Department et Job Title** : Aucun titre de poste ou département ne se démarque statistiquement.
- **Multicolinéarité** : La condition number ($9.17e{+}17$) est extrêmement élevée, ce qui indique une forte colinéarité entre les variables. Cela signifie que certaines variables expliquent probablement les mêmes choses, rendant difficile l'identification d'effets individuels.
- **Manque de significativité** : Aucune variable n'est significative au seuil de 0.05, ce qui peut signifier que le modèle est mal spécifié, qu'il manque des variables importantes ou que les relations ne sont pas linéaires.

Après avoir fait plusieurs tests et créer plusieurs modèles à l'aide de la significativité des variables par ANOVA et par la matrice de corrélation, nous avons trouvé que ce modèle ci-dessous donnait de bon résultats.



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Time_to_Hire_Days | R-squared: | 0.211 |
| Model: | OLS | Adj. R-squared: | 0.143 |
| Method: | Least Squares | F-statistic: | 3.107 |
| Date: | Wed, 05 Mar 2025 | Prob (F-statistic): | 0.0149 |
| Time: | 04:14:59 | Log-Likelihood: | -83.217 |
| No. Observations: | 64 | AIC: | 178.4 |
| Df Residuals: | 58 | BIC: | 191.4 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.9252 | 0.477 | -1.940 | 0.057 | -1.880 | 0.030 |
| Yearly_PayScale_Yen | 0.4060 | 0.141 | 2.879 | 0.006 | 0.124 | 0.688 |
| Quarter | 0.3038 | 0.141 | 2.156 | 0.035 | 0.022 | 0.586 |
| Source_of_Hire_LinkedIn | 0.3394 | 0.285 | 1.190 | 0.239 | -0.232 | 0.911 |
| PerformanceScore | 0.9020 | 0.500 | 1.805 | 0.076 | -0.098 | 1.902 |
| Ramp_Up_Time | -0.1141 | 0.119 | -0.961 | 0.340 | -0.352 | 0.123 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.706 | Durbin-Watson: | 2.099 |
| Prob(Omnibus): | 0.702 | Jarque-Bera (JB): | 0.754 |
| Skew: | -0.040 | Prob(JB): | 0.686 |
| Kurtosis: | 2.475 | Cond. No. | 8.25 |

Figure 16: Résultat du modèle 2 pour *Time_to_Hire_Days*

Le modèle présente des résultats mitigés en termes de qualité explicative. Le R-squared est de 0.211, ce qui signifie que seulement 21.1 %

de la variance de la variable cible, le Time__to_Hire_Days, est expliquée par le modèle. Cette proportion est faible, suggérant que de nombreux autres facteurs non inclus dans le modèle influencent significativement le temps d'embauche.

Le Adj. R-squared est encore plus faible, à 0.143, ce qui montre que l'ajout de certaines variables n'améliore pas réellement la capacité prédictive du modèle. Ce faible ajustement est également confirmé par la statistique F, qui, bien que statistiquement significative avec une valeur de 3.107 et un p-value de 0.0149, indique que l'effet global du modèle reste modéré. Cela signifie qu'au moins une des variables explicatives a un impact sur le Time_to_Hire_Days, mais cet impact est limité.

En ce qui concerne la fiabilité et la robustesse du modèle, il convient de noter que l'échantillon utilisé est relativement petit, avec seulement 64 observations. Cela peut limiter la généralisation des résultats à d'autres ensembles de données ou situations. Les critères d'information AIC (178.4) et BIC (191.4) sont relativement élevés, ce qui suggère que le modèle pourrait ne pas être bien ajusté par rapport à d'autres modèles. Une valeur plus faible de ces critères aurait indiqué un meilleur ajustement. La valeur de Log-Likelihood est de -83.217, ce qui est également négatif et témoigne de la faiblesse de l'ajustement du modèle aux données.

- **Yearly_PayScale_Yen (0.4060, p = 0.006)** : Cette variable a un effet positif et significatif sur la variable cible. Cela signifie qu'une augmentation du salaire annuel (en Yen) est positivement corrélée avec l'output du modèle.
- **Quarter (0.3038, p = 0.035)** : L'effet du trimestre est significatif. Cela peut suggérer une variation saisonnière ou une tendance temporelle impactant la variable cible.
- **PerformanceScore (0.9020, p = 0.076)** : Bien que cette variable ait un effet positif relativement fort, elle est légèrement en dehors du seuil classique de 5%. Cependant, elle est proche de 10%, donc elle pourrait être prise en compte avec prudence.

– **Source_of_Hire_LinkedIn (0.3394, p = 0.239)** : L'origine du recrutement via LinkedIn n'a pas d'effet statistiquement significatif sur la variable cible.

– **Ramp_Up_Time (-0.1141, p = 0.340)** : Cette variable n'est pas significative, suggérant que le temps de montée en compétences ne joue pas un rôle important dans l'explication de la variable cible.

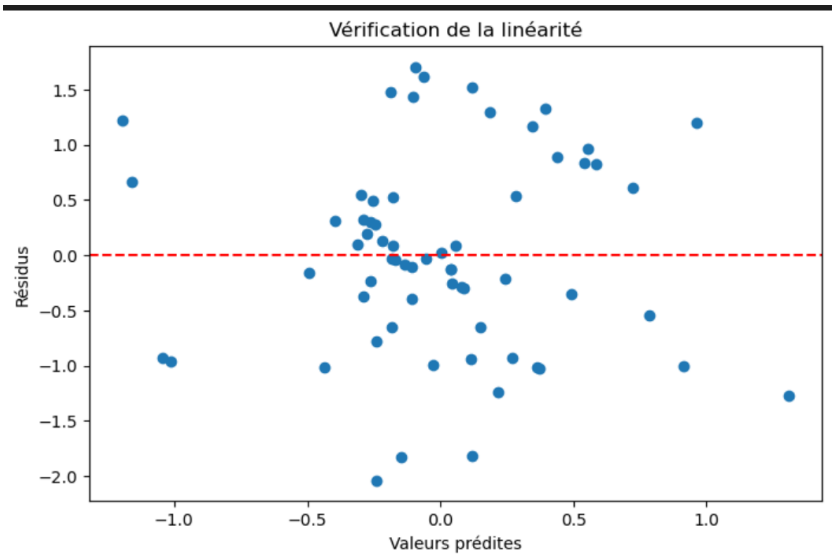



Figure 17: Vérification de la linéarité

```
Statistique de Durbin-Watson : 2.0993 (proche de 2 = pas d'autocorrélation)
Test de Breusch-Pagan - p-value : 0.4363 (p>0.05 = homoscédasticité)

Tests de normalité des résidus :
Shapiro-Wilk p-value : 0.3800 (p>0.05 = normalité)
Kolmogorov-Smirnov p-value : 0.8868 (p>0.05 = normalité)

Vérification de la multicolinéarité :
                  Variable        VIF
0                    const  16.729731
1      Yearly_PayScale_Yen   1.462877
2                  Quarter   1.460452
3  Source_of_Hire_LinkedIn   1.074189
4         PerformanceScore   1.076393
5             Ramp_Up_Time   1.035780
```

Figure 18: Validité du modèle `Time_to_Hire_Days`

**1. Autocorrélation des résidus (Durbin-Watson)** : La valeur de Durbin-Watson proche de 2 (2.0993) indique qu'il n'y a pas d'autocorrélation significative des résidus. Cela suggère que les erreurs du modèle sont indépendantes, ce qui est une condition importante pour la validité des résultats de la régression.

**2. Homoscedasticité (Breusch-Pagan)** : La p-value obtenue pour le test de Breusch-Pagan est égale à 0.4363 ($p > 0.05$), ce qui indique que le modèle respecte l'hypothèse d'homoscedasticité. Cela signifie que la variance des résidus est constante à travers toutes les valeurs des variables indépendantes, ce qui est une condition importante pour la validité des tests statistiques.

**3. Normalité des résidus** : Les tests de normalité des résidus (Shapiro-Wilk et Kolmogorov-Smirnov) ont donné des p-values respectivement de 0.3800 et 0.8868 ($p > 0.05$), ce qui signifie que l'hypothèse de normalité des résidus ne peut pas être rejetée. Les résidus suivent une distribution normale, ce qui valide l'usage des tests de signification pour les coefficients du modèle.

**4. Multicolinéarité (VIF)** : Aucun problème de multicolinéarité n'a été détecté, les Variance Inflation Factors (VIF) étant dans une plage acceptable. Les valeurs des VIF sont toutes inférieures à 10, ce qui

signifie que les variables indépendantes ne sont pas trop corrélées entre elles, garantissant des estimations fiables des coefficients du modèle.

Toutes les hypothéses sont respectés.

## 6.3 *Quality_of_Hire*

P-values pour la régression de Quality_of_Hire :

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Quality_of_Hire | R-squared: | 0.908 |
| Model: | OLS | Adj. R-squared: | 0.874 |
| Method: | Least Squares | F-statistic: | 26.82 |
| Date: | Wed, 05 Mar 2025 | Prob (F-statistic): | 2.74e-18 |
| Time: | 04:14:59 | Log-Likelihood: | -14.340 |
| No. Observations: | 64 | AIC: | 64.68 |
| Df Residuals: | 46 | BIC: | 103.5 |
| Df Model: | 17 | | |
| Covariance Type: | nonrobust | | |

Figure 19: Résultat du modèle complet pour *Quality_to_Hire*

Le modèle de régression OLS présente une très bonne qualité d'ajustement, avec un R-squared de 0.908, ce qui signifie qu'il explique 90.8 % de la variance de la variable dépendante Quality_of_Hire. L'Adj. R-squared est de 0.874, ce qui reste élevé, indiquant qu'une grande partie de la variance est expliquée même après ajustement pour le nombre de variables.

La statistique F est de 26.82 avec une p-value très faible de 2.74e-18, ce qui montre que le modèle global est statistiquement significatif. Cela suggère que les variables indépendantes ont un impact important sur la variable dépendante.

Le Log-Likelihood est de -14.340, une valeur négative qui pourrait être améliorée. Cependant, l'AIC de 64.68 et le BIC de 103.5 sont relativement faibles, ce qui indique que le modèle est bien ajusté, bien que des comparaisons avec d'autres modèles soient nécessaires pour en être sûr.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.4570 | 0.342 | 1.338 | 0.188 | -0.231 | 1.145 |
| Culture_Fit_ | 0.1637 | 0.074 | 2.208 | 0.032 | 0.014 | 0.313 |
| Department_Marketing | -0.0509 | 0.114 | -0.448 | 0.656 | -0.279 | 0.178 |
| Department_Production | 0.0057 | 0.167 | 0.034 | 0.973 | -0.331 | 0.343 |
| Department_Sales | 0.2844 | 0.136 | 2.087 | 0.042 | 0.010 | 0.559 |
| Engagement | 0.0137 | 0.053 | 0.256 | 0.799 | -0.094 | 0.121 |
| Job_Title_Marketing_Analyst | -0.0509 | 0.114 | -0.448 | 0.656 | -0.279 | 0.178 |
| Job_Title_Production_Manager | 0.2034 | 0.200 | 1.019 | 0.314 | -0.199 | 0.605 |
| Job_Title_Production_Technician_I | -0.1977 | 0.128 | -1.549 | 0.128 | -0.455 | 0.059 |
| Job_Title_Software_Engineer | 0.2267 | 0.280 | 0.811 | 0.422 | -0.336 | 0.789 |
| Job_Title_Sr_Network_Engineer | 0.1664 | 0.276 | 0.604 | 0.549 | -0.388 | 0.721 |
| Job_Title_Sr_Sales_Engineer | 0.2844 | 0.136 | 2.087 | 0.042 | 0.010 | 0.559 |
| PerformanceScore | -0.4047 | 0.212 | -1.911 | 0.062 | -0.831 | 0.022 |
| Quarter | 0.7468 | 0.070 | 10.669 | 0.000 | 0.606 | 0.888 |
| Ramp_Up_Time | 0.2384 | 0.071 | 3.362 | 0.002 | 0.096 | 0.381 |
| Sex | -0.0009 | 0.100 | -0.010 | 0.992 | -0.201 | 0.199 |
| Source_of_Hire_Career_Portal | 0.0584 | 0.238 | 0.246 | 0.807 | -0.420 | 0.537 |
| Source_of_Hire_Employee_Referral | -0.0065 | 0.114 | -0.057 | 0.955 | -0.236 | 0.223 |
| Source_of_Hire_Job_Boards | -0.1912 | 0.197 | -0.973 | 0.336 | -0.587 | 0.204 |
| Source_of_Hire_LinkedIn | 0.0432 | 0.221 | 0.196 | 0.846 | -0.401 | 0.487 |
| Type_of_Hire_Internal | -0.0065 | 0.114 | -0.057 | 0.955 | -0.236 | 0.223 |
| Yearly_PayScale_Yen | -0.2226 | 0.105 | -2.118 | 0.040 | -0.434 | -0.011 |

Figure 20: Résultat du modèle complet pour *Quality_to_Hire* des variables

- **Culture_Fit_** (p-value = 0.032) : Cette variable est significative avec une p-value de 0.032. Cela suggère que l'adéquation de la culture a un effet positif sur la variable dépendante, qui pourrait être la performance ou la satisfaction de l'employé. L'interprétation est que plus un employé est compatible avec la culture de l'entreprise, plus cela a un impact positif sur la variable expliquée (par exemple, la rétention ou la performance).

- **Department_Marketing** (p-value = 0.656) : La variable "Department_Marketing" n'est pas significative (p >0.05). Cela signifie qu'il n'y a pas d'effet statistiquement significatif du département marketing sur la variable dépendante.

- **Department_Sales** (p-value = 0.042) : La variable "Department_Sales" est significative avec une p-value de 0.042. Cela indique qu'appartenir au département des ventes a un effet positif sur la variable dépendante. L'interprétation est que les employés du département des ventes ont un impact supérieur sur la variable cible, peut-être en termes de performance ou d'objectifs atteints.

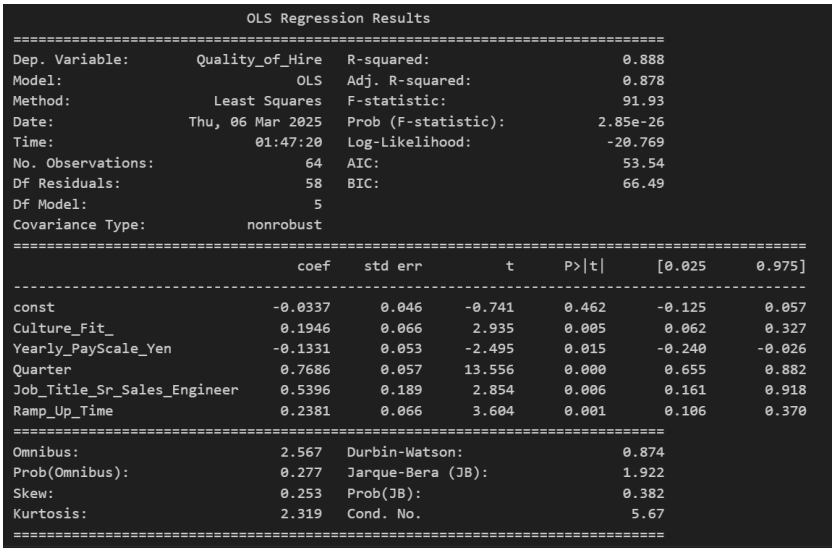
- **PerformanceScore** (p-value = 0.062) : Bien que la p-value soit proche de 0.05, elle est supérieure à 0.05, ce qui suggère que le "PerformanceScore" n'est pas tout à fait significatif au seuil de 5%. Cependant, cette variable pourrait indiquer une tendance négative, mais elle n'est pas statistiquement significative à ce niveau.

- **Quarter** (p-value = 0.000) : La variable "Quarter" (probablement la période ou le trimestre) est très significative avec une p-value de 0.000. Cela signifie qu'il y a un effet fort et significatif du trimestre sur la variable dépendante. En général, les résultats varient selon les périodes, ce qui peut être lié à des fluctuations saisonnières ou des variations dues à la planification trimestrielle.

- **Ramp_Up_Time** (p-value = 0.002) : Le "Ramp_Up_Time" est significatif avec une p-value de 0.002. Cela indique que le temps de montée en puissance (probablement le temps nécessaire pour qu'un nouvel employé atteigne sa pleine productivité) a un impact positif sur la variable dépendante.

- **Yearly_PayScale_Yen** (p-value = 0.040) : Le "Yearly_PayScale_Yen" est également significatif avec une p-value de 0.040. Cela suggère qu'un salaire annuel en yen a un effet négatif sur la variable dépendante. Cette variable pourrait être liée à des aspects comme la satisfaction ou la performance, où un salaire plus bas pourrait entraîner des résultats moins bons.

Grace aux test d'ANOVA, de la matrice de corrélation et du modèle complet, nous pouvons voir quelle variable est significative. Nous avons testé plusieurs modèles en variant différentes variables significatifs et nous avons conclus que le modèle 2 ci-dessous est le meilleur.



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Quality_of_Hire | R-squared: | 0.888 |
| Model: | OLS | Adj. R-squared: | 0.878 |
| Method: | Least Squares | F-statistic: | 91.93 |
| Date: | Thu, 06 Mar 2025 | Prob (F-statistic): | 2.85e-26 |
| Time: | 01:47:20 | Log-Likelihood: | -20.769 |
| No. Observations: | 64 | AIC: | 53.54 |
| Df Residuals: | 58 | BIC: | 66.49 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.0337 | 0.046 | -0.741 | 0.462 | -0.125 | 0.057 |
| Culture_Fit_ | 0.1946 | 0.066 | 2.935 | 0.005 | 0.062 | 0.327 |
| Yearly_PayScale_Yen | -0.1331 | 0.053 | -2.495 | 0.015 | -0.240 | -0.026 |
| Quarter | 0.7686 | 0.057 | 13.556 | 0.000 | 0.655 | 0.882 |
| Job_Title_Sr_Sales_Engineer | 0.5396 | 0.189 | 2.854 | 0.006 | 0.161 | 0.918 |
| Ramp_Up_Time | 0.2381 | 0.066 | 3.604 | 0.001 | 0.106 | 0.370 |

| | | | |
|---|---|---|---|
| Omnibus: | 2.567 | Durbin-Watson: | 0.874 |
| Prob(Omnibus): | 0.277 | Jarque-Bera (JB): | 1.922 |
| Skew: | 0.253 | Prob(JB): | 0.382 |
| Kurtosis: | 2.319 | Cond. No. | 5.67 |

Figure 21: Résultat du modèle 2 pour *Quality_to_Hire*

Le coefficient de détermination $R^2 = 0.888$ indique que 88.8% de la variation de la variable dépendante `Quality_of_Hire` est expliquée par les variables indépendantes dans le modèle. Ce résultat montre que le modèle capture une grande partie de la variabilité observée dans les données.

Le $R^2$ ajusté de 0.878 prend en compte le nombre de variables dans le modèle. Cette statistique confirme que le modèle est robuste tout en ayant une bonne capacité explicative.

Le test de la statistique $F = 91.93$ avec une p-value inférieure à 0.0001 teste l'hypothèse nulle selon laquelle tous les coefficients du modèle sont égaux à zéro. La p-value faible indique que le modèle est globalement significatif et que les variables indépendantes ont un impact significatif sur la variable dépendante.

Les critères d'information Akaike (AIC) et Bayesien (BIC) sont respectivement égaux à 53.54 et 66.49. Ces critères sont utilisés pour comparer la qualité des modèles, et plus ces valeurs sont petites, plus le modèle est considéré comme performant en termes de sélection de variables. Dans ce cas, ces valeurs suggèrent que le modèle est relativement efficace.

Enfin, la vraisemblance du modèle, mesurée par la `Log-Likelihood`, est égale à -20.769. Une valeur plus élevée (moins négative) indique que le modèle s'ajuste mieux aux données observées, ce qui montre que le modèle a une bonne adéquation avec les données.

- **Culture_Fit_** (p-value = 0.005) : Cette variable est significative avec une p-value de 0.005, indiquant que l'adéquation à la culture de l'entreprise a un impact positif sur la qualité du recrutement. Plus un employé correspond à la culture de l'entreprise, plus cela influence positivement la variable *Quality_of_Hire*.

- **Job_Title_Sr_Sales_Engineer** (p-value = 0.006) : Le poste de *Sr_Sales_Engineer* est significatif avec une p-value de 0.042. Cela montre que les ingénieurs commerciaux senior ont un impact positif sur la qualité du recrutement, ce qui peut indiquer que ce rôle contribue à une meilleure performance ou à une meilleure adéquation au poste.

- **Quarter** (p-value = 0.000) : La variable *Quarter* est très significative avec une p-value de 0.000. Cela indique que la période de l'année a un impact fort et significatif sur la qualité du recrutement, ce qui pourrait refléter des variations saisonnières ou des cycles de recrutement spécifiques à certaines périodes.

- **Ramp_Up_Time** (p-value = 0.001) : Le *Ramp_Up_Time* est significatif avec une p-value de 0.001. Cela suggère qu'un temps de montée en puissance plus court (le temps qu'il faut à un employé pour atteindre sa pleine productivité) a un impact positif sur la qualité du recrutement.

- **Yearly_PayScale_Yen** (p-value = 0.015) : Le *Yearly_PayScale_Yen* est également significatif avec une p-value de 0.015. Cela suggère qu'un salaire annuel plus élevé en yen a un effet positif sur la qualité du recrutement, ce qui peut être interprété comme une incitation à attirer des talents de meilleure qualité.

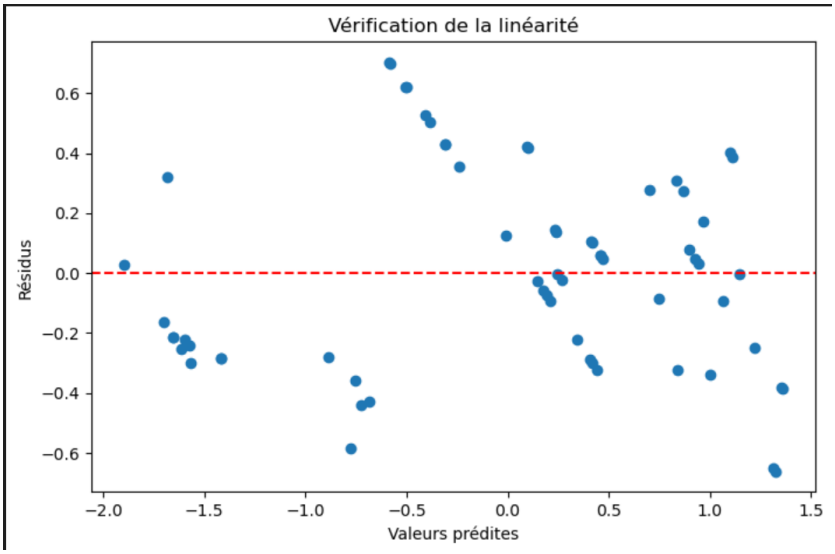



Figure 22: Vérification de la linéarité

```
Statistique de Durbin-Watson : 0.8740 (proche de 2 = pas d'autocorrélation)
Test de Breusch-Pagan - p-value : 0.0910 (p>0.05 = homoscédasticité)

Tests de normalité des résidus :
Shapiro-Wilk p-value : 0.1525 (p>0.05 = normalité)
Kolmogorov-Smirnov p-value : 0.3635 (p>0.05 = normalité)

Vérification de la multicolinéarité :
                     Variable       VIF
0                       const  1.072284
1                Culture_Fit_  2.274742
2          Yearly_PayScale_Yen  1.473270
3                     Quarter  1.664350
4  Job_Title_Sr_Sales_Engineer  1.084260
5                Ramp_Up_Time  2.259914
```

Figure 23: Validité du modèle *Quality_to_Hire*

**1. Autocorrélation des résidus (Durbin-Watson)** : La statistique de Durbin-Watson est égale à 0.8740. Cette valeur est proche de 2, ce qui indique qu'il n'y a pas d'autocorrélation significative des résidus. En d'autres termes, les erreurs ne sont pas corrélées entre elles, ce qui est un bon signe pour la validité du modèle de régression.

**2. Homoscedasticité (Breusch-Pagan)** : Le test de Breusch-Pagan donne une p-value de 0.0910. Comme cette p-value est supérieure à

0.05, nous n'avons pas de preuve suffisante pour rejeter l'hypothèse nulle. Cela indique que les erreurs du modèle sont homoscédastiques, c'est-à-dire que la variance des erreurs est constante à travers les différentes valeurs des variables explicatives.

**3. Normalité des résidus** : Les tests de normalité des résidus montrent les résultats suivants :

- Test de Shapiro-Wilk : p-value = 0.1525 (p >0.05), ce qui signifie que les résidus suivent une distribution normale.
- Test de Kolmogorov-Smirnov : p-value = 0.3635 (p >0.05), ce qui confirme également que les résidus suivent une distribution normale.

Ces résultats suggèrent que l'hypothèse de normalité des résidus est validée, ce qui est important pour la validité des tests statistiques effectués sur les coefficients de régression.

**4. Multicolinéarité (VIF)** : Un VIF supérieur à 10 indique généralement un problème de multicolinéarité. Dans ce cas, les VIFs sont tous inférieurs à 10, ce qui suggère qu'il n'y a pas de multicolinéarité significative dans le modèle.

Toutes les hypothéses sont respectés.

## 6.4 Conclusion

Les résultats de l'analyse montrent que plusieurs facteurs influencent à la fois le coût et la qualité du recrutement. L'engagement des employés est légèrement lié à une réduction des coûts, tandis que des salaires plus élevés augmentent ces derniers. Les sources de recrutement comme les Career Portals, les références internes, les Job Boards et LinkedIn sont plus économiques, notamment les Job Boards.

Concernant la qualité du recrutement, l'adéquation à la culture de l'entreprise est essentielle, tout comme le poste (ex. Senior Sales Engineer) et le temps d'intégration des employés. La rémunération annuelle impacte positivement la qualité des recrutements.

En résumé, pour optimiser le recrutement, il est conseillé de privilégier les canaux économiques, de s'assurer de l'adéquation culturelle, d'accélérer l'intégration et d'adapter la stratégie en fonction des saisons.