Projet de l'UE Statistique avec SAS

Kossi Tonyi Wobubey ABOTSI

Université de Strasbourg (2023-2024)

Table des matières

1 Introduction **3**

2 Exercice I **3**
- 2.1 Énoncé 3
- 2.2 Visualisation des données et formulation d'une hypothèse 3
- 2.3 Choix du test à réaliser 4
- 2.4 Vérification des conditions d'application 5
  - 2.4.1 Indépendance 5
  - 2.4.2 Exclusivité des classes 5
  - 2.4.3 Règle de Cochran 5
- 2.5 Conclusion 5

3 Exercice II **6**
- 3.1 Énoncé 6
- 3.2 Statistiques descriptives 6
- 3.3 Test statistique 6
- 3.4 Vérification des conditions d'application 7
  - 3.4.1 Indépendance 7
  - 3.4.2 Normalité 7
  - 3.4.3 Égalité des variances 7
- 3.5 Conclusion 8

4 Exercice III **8**
- 4.1 Énoncé 8
- 4.2 Statistique descriptive 9
  - 4.2.1 Tableau résumant les statistiques descriptives 9
  - 4.2.2 Visualisation des données 9
- 4.3 Modélisation 10

4.3.1 Élimination de l’information redondante dans notre modèle 13
4.3.2 Vérification de la surdispersion dans notre modèle 14
4.3.3 Sélection de modèle 14
4.4 Conclusion 20

1 Introduction

Ce projet, réalisé dans le cadre de l'unité d'enseignement "Statistiques avec SAS" au M1 Statistique, vise à appliquer les concepts statistiques à travers l'analyse de données réelles. Trois exercices explorent divers aspects de l'analyse statistique. Tous les tests statistiques effectués dans ce document sont réalisés avec un seuil $\alpha = 5\%$.

2 Exercice I

2.1 Énoncé

Nous disposons d'un échantillon de garçons et de filles provenant d'un district écossais. Cet échantillon est constitué d'observations indépendantes sur les caractéristiques de sexe et de couleur des cheveux, présentées dans un tableau.

Objectif : Analyser la relation entre la couleur des cheveux et le sexe.

2.2 Visualisation des données et formulation d'une hypothèse

Les données sont présentées sous forme de tableau avec les caractéristiques de sexe et de couleur des cheveux.

	BLOND	**ROUX**	**CHÂTAIN**	**BRUN**	**NOIR DE JAIS**
GARÇON	592	119	849	504	36
FILLE	544	97	677	451	14

TABLE 1 – Répartition des couleurs de cheveux par genre

Nous observons sur le graphique à barres empilées la répartition par sexe et couleur de cheveux (cf. script pour le code).

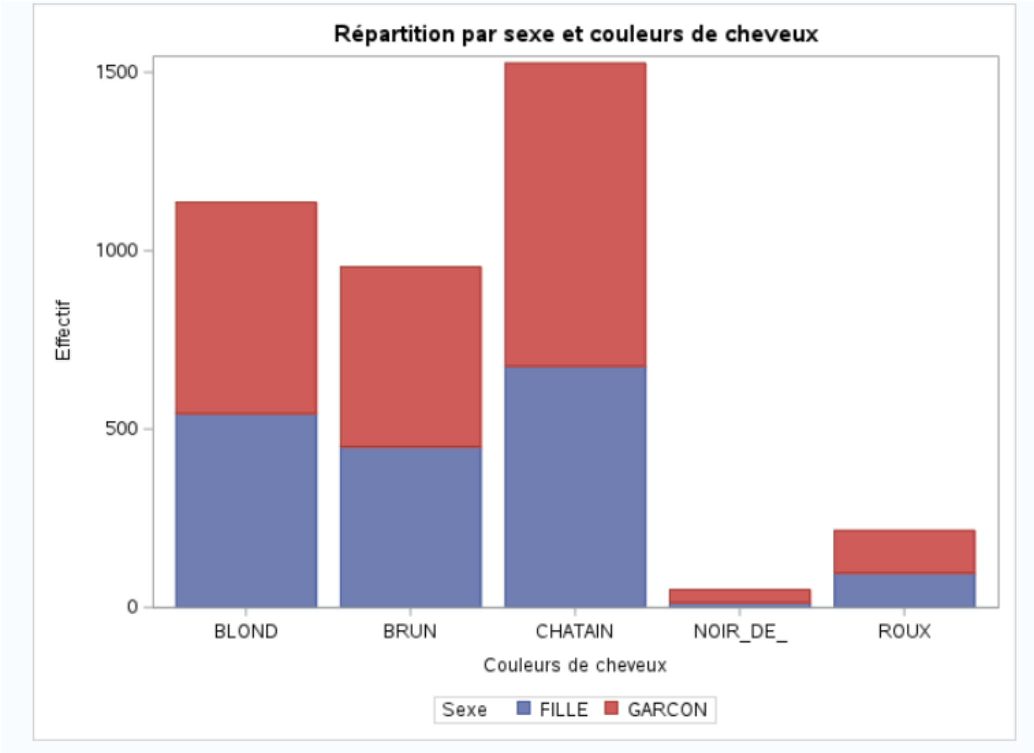

FIGURE 1 – Barres empilées de la répartition par sexe et couleur de cheveux

Ce graphique montre que les garçons ont légèrement plus de cheveux châtains et noirs de jais que les filles. Nous voulons donc déterminer si cette différence observée est statistiquement significative.

2.3 Choix du test à réaliser

Nous analysons deux variables qualitatives observées sur un échantillon pour tester leur indépendance. Le test le plus adapté dans cette situation est le **G-test**.

Les hypothèses à tester sont :

— H0 : Les variables sexe et couleur des cheveux sont indépendantes.

— H1 : Les variables sexe et couleur des cheveux ne sont pas indépendantes.

2.4 Vérification des conditions d'application

2.4.1 Indépendance

Les observations dans le tableau de contingence ont été collectées de manière indépendante.

2.4.2 Exclusivité des classes

Chaque individu observé est classé dans une et une seule catégorie pour chacune des deux variables.

2.4.3 Règle de Cochran

Pour respecter cette règle, au moins 80% des effectifs théoriques doivent être au moins égaux à 5.

Voici le tableau des effectifs théoriques obtenus :

	BLOND	**ROUX**	**CHÂTAIN**	**BRUN**	**NOIR DE JAIS**
GARÇON	614,37033	116,81689	825,28972	516,48210	27,04094
FILLE	521,62966	99,18310	700,71027	438,51789	22,95905

TABLE 2 – Répartition des couleurs de cheveux par genre avec valeurs décimales

Nous observons dans le tableau ci-dessus que plus de 80% des effectifs sont supérieurs à 5.

Toutes les conditions d'application étant vérifiées, nous pouvons procéder au G-test (cf. script SAS). Nous obtenons une statistique $\chi^2_{\text{obs}} = 10.7555$ qui, sous H0, suit une distribution χ^2_4 avec 4 degrés de liberté. La P-valeur observée est de 0.0295.

2.5 Conclusion

La P-valeur étant inférieure à 5%, nous rejetons H0. Par conséquent, les variables sexe et couleur de cheveux ne sont pas indépendantes, ce qui indique que la différence observée est statistiquement significative. Le V de Cramer nous donne 0.0519, inférieur à 0.1, donc l'intensité de cette liaison est très faible mais pas négligeable.

3 Exercice II

3.1 Énoncé

Une entreprise pharmaceutique expérimente une nouvelle molécule censée faire baisser une mesure physiologique chez des patients porteurs d'une pathologie. Pour cela, 54 patients sans lien de parenté ont été recrutés. Chaque patient s'est vu attribuer soit la nouvelle molécule (traitement A) soit un placebo (traitement B). Les groupes ont été constitués de manière à ce que les distributions de l'âge et du sexe soient similaires dans les deux groupes. Les données sont contenues dans le fichier `molecule.csv`.

Questions :

1. Quelles sont les moyennes et écarts-types dans chacun des groupes ?
2. Est-ce que la molécule testée est efficace ?

Objectif : Tester l'efficacité de la nouvelle molécule.

3.2 Statistiques descriptives

Les statistiques descriptives de notre jeu de données, où les valeurs affichées correspondent à une mesure physiologique pour des patients porteurs d'une pathologie, sont les suivantes :

	Traitement A	Traitement B
Nombre d'observations	27	27
Moyenne	78.836	65.751
Écart-type	11.141	12.416

TABLE 3 – Moyenne et Écart-type dans chacun des groupes

On observe que en moyenne la mesure physiologique des patients ayant subi le traitement A est supérieure à celle des patients ayant reçu le traitement B. La question maintenant est de savoir si cette différence observée est significative.

3.3 Test statistique

Pour déterminer si la molécule testée est efficace, nous allons comparer les moyennes des mesures effectuées sur les deux groupes. Ayant seulement deux groupes et une variable quantitative, le test adéquat à réaliser est le test de **Student**.

Les hypothèses à tester sont :

— H0 : La mesure moyenne des patients ayant reçu le traitement A est la même que celle des patients ayant reçu le traitement B (placebo).
— H1 : La mesure moyenne des patients ayant reçu le traitement A est différente de celle des patients ayant reçu le traitement B (placebo).

3.4 Vérification des conditions d'application

3.4.1 Indépendance

Selon le protocole de collecte des données, les patients sans lien de parenté ont été recrutés, ce qui permet de conclure que nos observations sont indépendantes.

3.4.2 Normalité

Soit Y notre variable d'intérêt et Y_1, Y_2 respectivement les variables dans les échantillons de **Traitement A** et **Traitement B**. Nous allons procéder au test de **Shapiro-Wilk** sur nos deux échantillons.

Les hypothèses à tester sont :
— H0 : L'échantillon est issu d'une population normalement distribuée.
— H1 : L'échantillon n'est pas issu d'une population normalement distribuée.

Nos deux échantillons sont composés d'observations indépendantes issues d'une variable d'intérêt qui est quantitative continue. La condition d'application du test de **Shapiro-Wilk** est donc vérifiée.

Normalité de Y_1 : En procédant au test (cf. script), on obtient une p-valeur de **0.8149**. La p-valeur observée est supérieure au seuil α, donc nous conservons l'hypothèse nulle H0. L'échantillon est issu d'une population normalement distribuée.

Normalité de Y_2 : En procédant au test (cf. script), on obtient une p-valeur de **0.9158**. La p-valeur observée est supérieure au seuil α, donc nous conservons l'hypothèse nulle H0. L'échantillon est issu d'une population normalement distribuée.

3.4.3 Égalité des variances

Nous allons vérifier l'égalité des variances dans nos deux échantillons. Soit σ_1^2 et σ_2^2 les variances respectives inconnues des échantillons **Traitement A** et **Traitement B**. Pour ce faire, nous allons procéder au test de **Fisher-Snedecor**. Les hypothèses sont :
— H0 : $\sigma_1^2 = \sigma_2^2$
— H1 : $\sigma_1^2 \neq \sigma_2^2$

Les conditions d'application sont :

— Chaque échantillon est composé d'observations indépendantes.
— Les deux échantillons sont indépendants.
— $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$
— $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$

Les deux premières conditions sont déjà vérifiées, et les deux dernières sont vérifiées par les tests de normalité précédents. En procédant au test (cf. script), on obtient : La statistique de décision **F = 1.24**, qui sous H0 suit une loi de Fisher à 26 et 26 degrés de liberté. La p-valeur est de **0.5846**. La p-valeur est supérieure au seuil α, donc nous conservons H0 : il y a égalité des variances.

3.5 Conclusion

Toutes les conditions d'application du test de Student sont vérifiées. Nous réalisons donc le test (cf. script). Nous observons une statistique **T = 4.08**, qui sous H0 suit la loi de Student à 52 degrés de liberté. La p-valeur est de **0.0002**. La p-valeur étant inférieure au seuil α, nous rejetons H0. Donc, la moyenne des mesures effectuées dans les deux groupes est différente. On observe aussi que la moyenne de la mesure du groupe ayant eu le traitement A est supérieure à celle ayant eu le traitement B, ce qui indique que la molécule testée est inefficace puisqu'elle était censée faire baisser la mesure physiologique. Elle a plutôt un effet inverse.

4 Exercice III

4.1 Énoncé

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme l'ozone (O3). Le jeu de données `ozone.xls` comporte 112 observations indépendantes relevées durant l'été 2001.

— La variable maxO3 est le maximum journalier de la concentration en ozone (en $\mu g/m^3$).
— Les variables T9, T12, T15 correspondent aux températures relevées respectivement à 9h, 12h et 15h.
— Les variables Ne9, Ne12, Ne15 correspondent aux nébulosités relevées respectivement à 9h, 12h et 15h.
— Les variables Vx9, Vx12, Vx15 correspondent à la composante Est-Ouest du vent relevée respectivement à 9h, 12h et 15h.

Le but de l'exercice est d'étudier l'éventuelle relation linéaire entre la variable maxO3 et l'ensemble des autres variables. Pour cela, à l'aide de SAS, après avoir présenté les statistiques descriptives des variables, il faut appliquer un protocole de sélection du modèle. Le critère de sélection utilisé devra être présenté. Une fois le modèle optimal trouvé, il est demandé d'interpréter les résultats.

Objectif : Étudier la relation linéaire entre la variable maxO3 et l'ensemble des autres variables.

4.2 Statistique descriptive

4.2.1 Tableau résumant les statistiques descriptives

(cf. script pour le code)

Variable	N	Nbre manquant	Moyenne	Écart-type	Min	1er Quartile	Médiane	3ème Quartile	Max
maxO3	112	0	90.3035714	28.1872245	42	70.5	81.5	106	166
T9	112	0	18.3607143	3.1227257	11.3	16.2	17.8	19.95	27
T12	112	0	21.5267857	4.0423208	14	18.6	20.55	23.6	33.5
T15	112	0	22.6276786	4.5308594	14.9	19.25	22.05	25.6	35.5
Ne9	112	0	4.9285714	2.5949163	0	3	6	7	8
Ne12	112	0	5.0178571	2.2818601	0	4	5	7	8
Ne15	112	0	4.8303571	2.3322587	0	3	5	7	8
Vx9	112	0	-1.2143455	2.6327423	-7.8785	-3.339	-0.866	0.6946	5.1962
Vx12	112	0	-1.6110036	2.7956729	-7.8785	-3.6294	-1.8794	0	6.5778
Vx15	112	0	-1.690683	2.8101977	-9	-3.9392	-1.54965	0	5

TABLE 4 – Statistiques descriptives

4.2.2 Visualisation des données

Observons les nuages de points par paires (cf. script pour le code).

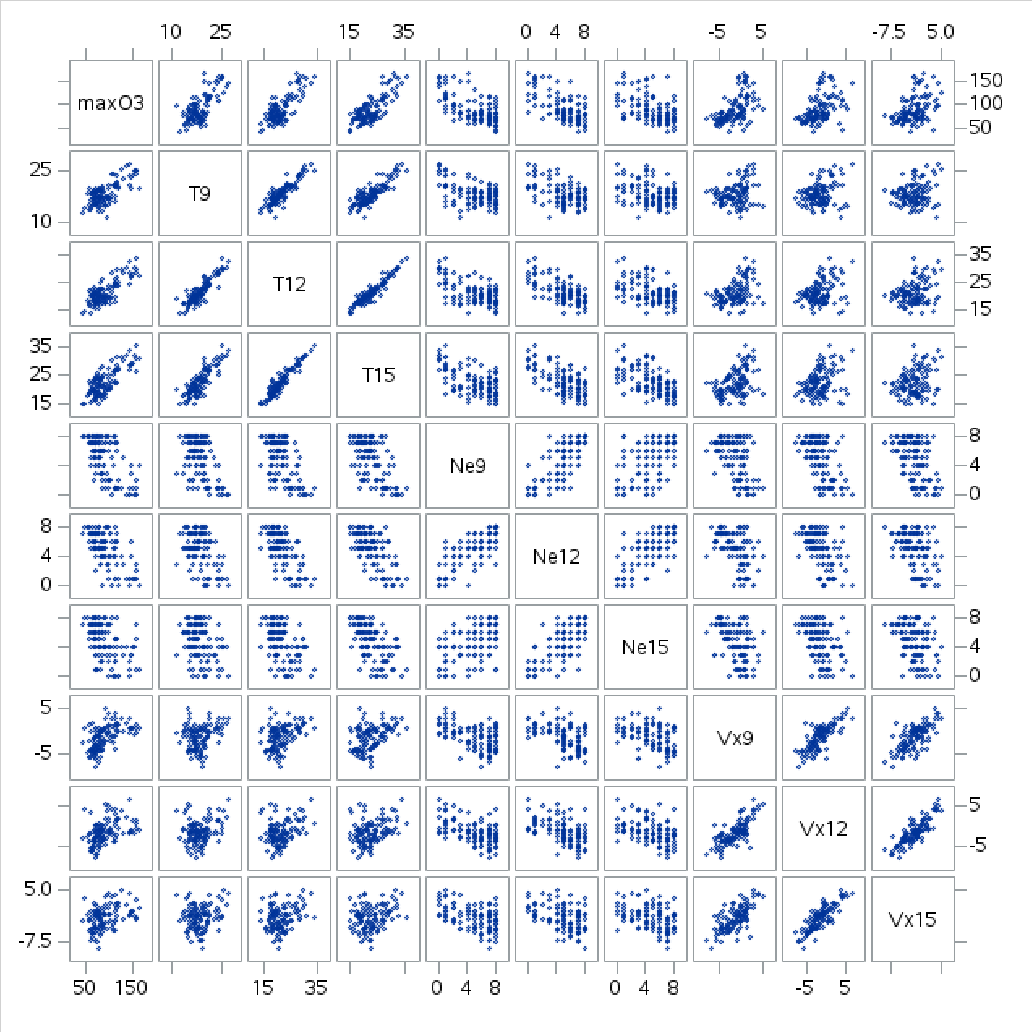

FIGURE 2 – Nuage de points

Comme on peut le voir ci-dessus, on remarque une relation linéaire (forte corrélation) entre les variables T9, T12, T15 et entre les variables Vx9, Vx12, Vx15.

4.3 Modélisation

Nous allons étudier la régression de la variable maxO3 par les variables T9, T12, T15, Ne9, Ne12, Ne15, Vx9, Vx12, Vx15 dans le cadre d'un modèle linéaire généralisé (GLM) de Poisson.

La variable maxO3 est supposée suivre une loi de Poisson de paramètre λ car discrète et positive. Nos observations étant indépendantes grâce au protocole de collecte des données, nous utiliserons un GLM de Poisson pour étudier la relation linéaire entre maxO3 et les variables explicatives. La fonction de lien pour ce modèle est la fonction log.

Soit Y = maxO3. Nous avons donc $log(E(Y_i)) = \eta_i$, avec η_i une combinaison linéaire des variables explicatives (T9, T12, T15, Ne9, Ne12, Ne15, Vx9, Vx12, Vx15), combinaison dont nous cherchons les coefficients, plus un coefficient constant.

Tous les résultats présentés ont leurs codes dans le script SAS.

Voici un extrait de la sortie liée à notre modèle :

Nombre d'observations lues	112
Nombre d'observations utilisées	112

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Ecart	102	281.8196	2.7629
Déviance normalisée	102	281.8196	2.7629
Khi2 de Pearson	102	284.7721	2.7919
Pearson normalisé X2	102	284.7721	2.7919
Log-vraisemblance		35748.3946	
Log-vraisemblance complète		-493.6561	
AIC (préférer les petites valeurs)		1007.3122	
AICC (préférer les petites valeurs)		1009.4905	
BIC (préférer les petites valeurs)		1034.4972	

L'algorithme a convergé.

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	3.7404	0.0956	3.5530	3.9277	1530.59	<.0001
T9	1	0.0170	0.0078	0.0017	0.0324	4.71	0.0299
T12	1	0.0141	0.0106	-0.0067	0.0348	1.77	0.1837
T15	1	0.0122	0.0085	-0.0044	0.0289	2.08	0.1491
Ne9	1	-0.0119	0.0066	-0.0249	0.0011	3.22	0.0728
Ne12	1	-0.0130	0.0094	-0.0315	0.0055	1.90	0.1678
Ne15	1	-0.0021	0.0072	-0.0162	0.0119	0.09	0.7640
Vx9	1	0.0285	0.0069	0.0151	0.0420	17.32	<.0001
Vx12	1	-0.0093	0.0076	-0.0243	0.0056	1.49	0.2223
Vx15	1	0.0036	0.0067	-0.0096	0.0168	0.28	0.5948
Echelle	0	1.0000	0.0000	1.0000	1.0000		

Note: Le paramètre d'échelle a été fixé.

FIGURE 3 – Résultats du GLM de Poisson

Ci-dessus nous avons l'estimation des paramètres du modèle et les tests de validité de ces paramètres. Soit β_j un paramètre du modèle sauf la constante. Les hypothèses du test de validité de ce paramètre sont :

— H0 : $\beta_j = 0$

— H1 : $\beta_j \neq 0$

On utilise la statistique de test Q = $\hat{\beta_j}^2 / \hat{\sigma_j}^2$ qui, sous H0, suit une loi du χ^2 à 1 ddl. Pour un niveau $\alpha = 5\%$, on acceptera la nullité des paramètres associés aux variables T12, T15, Ne9, Ne12, Ne15, Vx12 et Vx15 car leurs p-valeurs observées sont supérieures à 5% et on rejettera la nullité des paramètres associés aux variables T9 et Vx9 car leurs p-valeurs sont inférieures à 5%.

4.3.1 Élimination de l'information redondante dans notre modèle

Nous allons étudier la corrélation entre les variables explicatives. Avant de procéder à la sélection de modèle, il est crucial de retirer toute redondance parmi les variables explicatives. Idéalement, ces variables devraient être indépendantes les unes des autres pour éviter des problèmes de multicolinéarité. Afin d'identifier et d'éliminer toute corrélation excessive entre elles, l'utilisation du Variance Inflation Factor (VIF) est recommandée. Un seuil de VIF = 10 est utilisé car notre modèle n'a pas de problème de sur-paramétrage, avec 112 observations pour au plus 10 paramètres.

On obtient le résultat suivant :

Variable	**VIF**
Intercept	0
T9	6.02420
T12	17.99033
T15	14.44099
Ne9	3.08475
Ne12	5.18893
Ne15	2.93896
Vx9	2.94865
Vx12	4.65468
Vx15	3.56281

TABLE 5 – Variables et VIF correspondants

On remarque que le VIF des variables explicatives T12 et T15 est supérieur à 10. Nous éliminons donc celle ayant le plus grand VIF, T12, et répétons l'analyse. On obtient :

Variable	VIF
Intercept	0
T9	4.57296
T15	6.73019
Ne9	3.03129
Ne12	3.99759
Ne15	2.43735
Vx9	2.85535
Vx12	4.47157
Vx15	3.51988

TABLE 6 – Variables et VIF correspondants après élimination

Tous les VIF sont maintenant inférieurs à 10. Nous avons donc notre modèle optimal au sens du VIF, c'est-à-dire que les variables explicatives ne sont pas trop corrélées.

4.3.2 Vérification de la surdispersion dans notre modèle

La variable maxO3 modélisée par une loi de Poisson mais la variance n'est pas forcément égale à l'espérance. Dans notre cas, la variance est $\phi\times$espérance, l'idéal étant que $\phi = 1$ avec $\phi =$ déviance$/(n - p)$, où p est le nombre de paramètres et n l'effectif de l'échantillon. Or, on observe $\phi = 2.7629$, supérieur à 1, ce qui indique une surdispersion.

Pour corriger ce problème, nous introduisons un paramètre dans le modèle permettant de ramener ϕ à 1. Le code incluant la correction se trouve dans le script SAS. Nous effectuons ensuite la sélection de modèle avec cette correction.

4.3.3 Sélection de modèle

Nous allons maintenant entreprendre la sélection du modèle. Notre modèle n'a pas de problème de surparamétrage. L'idéal serait de faire une sélection de modèle à l'aide de l'AIC avec la méthode du backward. Cependant, avec la correction effectuée due à la surdispersion, nous ne pouvons plus calculer la vraisemblance car la loi n'est plus Poisson mais quasi-Poisson. Donc le choix qui s'offre à nous est les tests LRT.

Hypothèses :

— H0 : Il n'y a pas de différence significative entre M1 et M2 (M1 : modèle complet, M2 : modèle avec une variable de moins)

— H1 : Il y a une différence significative entre M1 et M2

En effectuant le test, voici le premier résultat :

Statistique LR pour Analyse de Type 3						
Source	DDL num.	DDL den.	Valeur F	Pr > F	Khi-2	Pr > khi-2
T9	1	103	3.90	0.0511	3.90	0.0484
T15	1	103	4.83	0.0301	4.83	0.0279
Ne9	1	103	0.98	0.3256	0.98	0.3233
Ne12	1	103	1.83	0.1790	1.83	0.1760
Ne15	1	103	0.03	0.8719	0.03	0.8716
Vx9	1	103	7.20	0.0085	7.20	0.0073
Vx12	1	103	0.76	0.3867	0.76	0.3847
Vx15	1	103	0.14	0.7050	0.14	0.7042

FIGURE 4 – Résultats du test LRT

La variable Ne15 a une p-valeur de **0.8716**, la plus grande, supérieure au seuil de 5%, donc si elle est éliminée du modèle, il n'y a aucune différence significative avec l'ancien modèle.

Effectuons de nouveau le test sans la variable Ne15, on obtient :

Statistique LR pour Analyse de Type 3						
Source	DDL num.	DDL den.	Valeur F	Pr > F	Khi-2	Pr > khi-2
T9	1	104	4.64	0.0335	4.64	0.0312
T15	1	104	5.37	0.0225	5.37	0.0205
Ne9	1	104	1.00	0.3192	1.00	0.3169
Ne12	1	104	2.01	0.1588	2.01	0.1558
Vx9	1	104	7.50	0.0073	7.50	0.0062
Vx12	1	104	0.79	0.3751	0.79	0.3731
Vx15	1	104	0.14	0.7103	0.14	0.7096

FIGURE 5 – Résultats du test LRT

La variable Vx15 a une p-valeur de **0.7096**, la plus grande, supérieure au seuil de 5%, donc si elle est éliminée du modèle, il n'y a aucune différence significative avec l'ancien modèle.

Effectuons de nouveau le test sans la variable Vx15, on obtient :

Statistique LR pour Analyse de Type 3						
Source	DDL num.	DDL den.	Valeur F	Pr > F	Khi-2	Pr > khi-2
T9	1	105	4.62	0.0338	4.62	0.0315
T15	1	105	5.43	0.0217	5.43	0.0198
Ne9	1	105	1.18	0.2808	1.18	0.2783
Ne12	1	105	1.94	0.1664	1.94	0.1635
Vx9	1	105	8.13	0.0053	8.13	0.0044
Vx12	1	105	0.72	0.3992	0.72	0.3973

FIGURE 6 – Résultats du test LRT

La variable Vx12 a une p-valeur de **0.3973**, la plus grande, supérieure au seuil de 5%, donc si elle est éliminée du modèle, il n'y a aucune différence significative avec l'ancien modèle.

Effectuons de nouveau le test sans la variable Vx12, on obtient :

Statistique LR pour Analyse de Type 3						
Source	DDL num.	DDL den.	Valeur F	Pr > F	Khi-2	Pr > khi-2
T9	1	106	4.23	0.0422	4.23	0.0398
T15	1	106	6.47	0.0124	6.47	0.0110
Ne9	1	106	0.97	0.3265	0.97	0.3242
Ne12	1	106	1.63	0.2039	1.63	0.2011
Vx9	1	106	9.70	0.0024	9.70	0.0018

FIGURE 7 – Résultats du test LRT

La variable Ne9 a une p-valeur de **0.3242**, la plus grande, supérieure au seuil de 5%, donc si elle est éliminée du modèle, il n'y a aucune différence significative avec l'ancien modèle.

Effectuons de nouveau le test sans la variable Ne9, on obtient :

Statistique LR pour Analyse de Type 3						
Source	DDL num.	DDL den.	Valeur F	Pr > F	Khi-2	Pr > khi-2
T9	1	107	5.39	0.0221	5.39	0.0202
T15	1	107	6.11	0.0150	6.11	0.0135
Ne12	1	107	6.02	0.0158	6.02	0.0142
Vx9	1	107	11.87	0.0008	11.87	0.0006

FIGURE 8 – Résultats du test LRT

Toutes les variables (T9, T15, Ne12, Vx9) ont une p-valeur inférieure au seuil de 5%, donc si l'une d'entre elles est éliminée, il y aura une différence significative avec l'ancien modèle.

Voici le résumé du modèle final, le modèle optimal après la sélection de modèle :

Nombre d'observations lues	112
Nombre d'observations utilisées	112

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Ecart	107	288.6005	2.6972
Déviance normalisée	107	107.0000	1.0000
Khi2 de Pearson	107	290.2372	2.7125
Pearson normalisé X2	107	107.6068	1.0057
Log-vraisemblance		13252.6288	
Log-vraisemblance complète		-497.0466	
AIC (préférer les petites valeurs)		1004.0932	
AICC (préférer les petites valeurs)		1004.6592	
BIC (préférer les petites valeurs)		1017.6857	

L'algorithme a convergé.

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance de Wald à95%		Khi-2 de Wald	Pr > khi-2
Intercept	1	3.7188	0.1448	3.4350	4.0026	659.50	<.0001
T9	1	0.0237	0.0102	0.0037	0.0436	5.40	0.0201
T15	1	0.0208	0.0084	0.0043	0.0373	6.11	0.0135
Ne12	1	-0.0234	0.0095	-0.0421	-0.0047	6.00	0.0143
Vx9	1	0.0266	0.0077	0.0115	0.0417	11.91	0.0006
Echelle	0	1.6423	0.0000	1.6423	1.6423		

Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

FIGURE 9 – Modèle final

Interprétation :

Tous les paramètres du modèle (Intercept, T9, T15, Ne12, Vx9) sont statistiquement significatifs, avec des p-valeurs inférieures à 0.05, ce qui signifie que chacun de ces paramètres a un effet significatif sur la variable de réponse dans le modèle de Poisson.

- **Intercept**
  - **Estimation** : 3.7188
  - **Erreur type** : 0.1448
  - **Intervalle de confiance de Wald à 95%** : [3.4350, 4.0026]
  - **Khi-2 de Wald** : 659.50
  - **Pr > khi-2** : <0.0001

L'intercept représente le log du maxO3 lorsque toutes les variables indépendantes sont égales à zéro. Son estimation est très significative (p-valeur < 0.0001), indiquant qu'il y a un effet de base substantiel.

- **T9**
  - **Estimation** : 0.0237
  - **Erreur type** : 0.0102
  - **Intervalle de confiance de Wald à 95%** : [0.0037, 0.0436]
  - **Khi-2 de Wald** : 5.45
  - **Pr > khi-2** : 0.0195

  La variable T9 a un coefficient positif et significatif (p-valeur = 0.0195). Cela signifie que pour chaque unité d'augmentation de T9, le log du maxO3 augmente de 0.0237, toute chose étant égale par ailleurs, ce qui correspond à une valeur de maxO3 multipliée par $\exp(0.0237) \approx 1.024$.
- **T15**
  - **Estimation** : 0.0208
  - **Erreur type** : 0.0084
  - **Intervalle de confiance de Wald à 95%** : [0.0043, 0.0373]
  - **Khi-2 de Wald** : 6.16
  - **Pr > khi-2** : 0.0131

  La variable T15 a également un coefficient positif et significatif (p-valeur = 0.0131). Une augmentation d'une unité de T15 entraîne une augmentation du log du maxO3 de 0.0208, toute chose étant égale par ailleurs, soit une valeur de maxO3 multipliée par $\exp(0.0208) \approx 1.021$.
- **Ne12**
  - **Estimation** : -0.0234
  - **Erreur type** : 0.0095
  - **Intervalle de confiance de Wald à 95%** : [-0.0420, -0.0047]
  - **Khi-2 de Wald** : 6.02
  - **Pr > khi-2** : 0.0142

  La variable Ne12 a un coefficient négatif et significatif (p-valeur = 0.0142). Cela signifie qu'une augmentation d'une unité de Ne12 entraîne une diminution du log de maxO3 de 0.0234, toute chose étant égale par ailleurs, soit une valeur de maxO3 multipliée par $\exp(-0.0234) \approx 0.977$.
- **Vx9**
  - **Estimation** : 0.0206
  - **Erreur type** : 0.0077
  - **Intervalle de confiance de Wald à 95%** : [0.0115, 0.0297]
  - **Khi-2 de Wald** : 11.91
  - **Pr > khi-2** : 0.0006

  La variable Vx9 a un coefficient positif et très significatif (p-valeur = 0.0006). Une augmentation d'une unité de Vx9 augmente le log du maxO3 de 0.0206, toute chose étant égale par ailleurs, soit une valeur de maxO3 multipliée par $\exp(0.0206) \approx 1.021$.

4.4 Conclusion

Les variables T9, T15, et Vx9 ont des effets positifs et significatifs sur la variable maxO3, tandis que Ne12 a un effet négatif et significatif. Ces résultats peuvent être utilisés pour comprendre les influences des différentes variables sur la variable dépendante, en tenant compte des effets multiplicatifs sur le taux d'incidence dans un cadre de modèle de Poisson.