

Guide d'utilisation du modèle « In-linéaire généralisé », outil développé pour le traitement des données du suivi *in situ* des installations d'ANC (Boutin et *al.*, 2017) en collaboration avec Yves Le Gat (INRAE Nouvelle-Aquitaine Bordeaux)

Sommaire

➤ Création du modèle « In-linéaire généralisé » adapté à l'existence de données censurées	2
➤ Exemple de l'application du modèle « In-linéaire généralisé » aux données du suivi <i>in situ</i>	3
➤ Utilisation pratique à partir d'un autre cas d'étude que le suivi <i>in situ</i>	6
Script de base à utiliser	6
Données d'entrée à fournir au script.....	6
Utilisation du script	6
Point de vigilance	10
Interprétation des résultats	10
Autre exemple avec des effets conjoints, c'est à dire plusieurs effets simultanés identifiés	12
➤ Utilisation de la version développée sous R-Shiny.....	13

Bibliographie

BOUTIN C., OLIVIER L., AGENET Ph., PARISI S., ARTUIT P., BRANCHU Ph., DECOUT A., DUBOIS V., 818 DUBOURG L., DHUMEAUX D., JOUSSE S., LEVAL C., MOULINE B., PORTIER N., RAMBERT C., SOULIAC 819 L. et SZABO C. (2017) : « Assainissement non collectif : Le suivi *in situ* des installations de 2011 à 2016 » Rapport final 186 p., disponible en ligne : [hal-02606167v1](https://hal.archives-ouvertes.fr/hal-02606167v1)

HILL Catherine (1999) : Analyse statistique des données de survie, Médecine-Sciences, Paris, ISBN 2257123107, 190 p.

NELDER J. A., WEDDERBURN R. W. M. (1972) : « Modèles linéaires généralisés. » *Journal of the Royal Statistical Society*; 135, 370-384.

➤ Création du modèle « In-linéaire généralisé » adapté à l'existence de données censurées

Le modèle « In-linéaire généralisé » a été créé pour répondre à la présence de données censurées dans un jeu de données à étudier, par exemple du fait des limites de quantification (LQ) des appareils de mesure qui indiquent le seuil en dessous duquel la grandeur mesurée n'est plus quantifiable.

La censure à gauche, dans le langage des statisticiens, est un concept théorique permettant de transformer les données basses tout en les maintenant dans le jeu de données. Les outils statistiques utilisés pour l'analyse des données censurées ont été développés, en grande majorité, pour la médecine et l'épidémiologie (Catherine Hill, 2009). Cependant, les concepts et méthodes ont une portée générale et peuvent être utilisés dans d'autres domaines.

Il existe trois types de censures : la censure à gauche, la censure à droite et la censure par intervalle. Les données sont dites « censurées à gauche » lorsqu'elles se trouvent en dessous d'un seuil défini. Les données « censurées à droite » sont supérieures à un seuil fixé, et les données « censurées par intervalle » se trouvent comprises entre une borne inférieure et une borne supérieure.

Ici, les données « **censurées à gauche** » sont celles inférieures à la LQ, et toutes les données inférieures à la LQ sont remplacées par un intervalle : $[0 ; LQ]$.

Les modèles linéaires généralisés (GLiM) ont été introduits par NELDER et WEDDERBURN [1972] comme un outil probabiliste flexible permettant de modéliser la distribution des données observées en fonction de facteurs explicatifs.

Ici, le modèle suivant est considéré, dans lequel la variable dépendante est transformée en logarithme népérien afin d'assurer la positivité des prédictions :

$$\ln(Y) = \mu + s * \omega$$

où :

- Y représente la variable aléatoire dépendante ;
- $\mu = X^T \beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ est la combinaison linéaire des variables explicatives et de leurs coefficients de régression ;
- $X = (1 \ X_1 \ X_2 \ \dots \ X_p)^T$ est le vecteur des p variables explicatives, considérées comme non aléatoires ;
- $\beta = (\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_p)^T$ est le vecteur des coefficients de régression qui quantifient les effets des variables explicatives sur la médiane de $\ln(Y)$;
- s est le paramètre de dispersion du modèle ;
- ω est une variable aléatoire normale $N(0,1)$.

Les coefficients $(\beta_1, \beta_2, \dots)$ sont les effets des variables explicatives (X_1, X_2, X_3, \dots) sur la médiane de la variable dépendante (Y) . Le calage du modèle In-linéaire consiste à chercher les coefficients $(\beta_1, \beta_2, \dots ; s)$ qui maximisent la fonction de vraisemblance, i.e. le plus vraisemblable au regard des observations, en réduisant s .

Une fois calés, les coefficients $(\beta_1, \beta_2, \dots)$ estimés par le modèle permettent de calculer les médianes. Ils permettent aussi de calculer des facteurs d'accroissement ou de diminution $(e^{\beta_1}, e^{\beta_2}, \dots)$ exprimés en pourcentage vis-à-vis de l'ensemble de référence (expliqué ci-après).

Une fonctionnalité supplémentaire du modèle consiste à faire en sorte que le paramètre de **dispersion** s dépende de q variables explicatives, au lieu d'être constant pour toutes les observations. Le modèle de dispersion choisi est $s = e^{(\gamma_0 + \gamma_1 X_1 + \dots + \gamma_q X_q)}$ dans lequel la transformation exponentielle assure la positivité de s et dans lequel γ représente les coefficients de régression des variables explicatives sur les coefficients de dispersion. Par rapport à un modèle sans dispersion, le fait de tenir compte de la dispersion modifie l'estimation des coefficients β des variables explicatives (X) et améliore la précision de l'ajustement du modèle.

C'est ce modèle In-linéaire généralisé qui a été appliqué au jeu de données du suivi *in situ*.

➤ Exemple de l'application du modèle « In-linéaire généralisé » aux données du suivi *in situ*

Les modèles « In-linéaire généralisés » permettent de comparer un ensemble de données (Y) en identifiant les effets des différentes variables explicatives. Les variables explicatives (X) sont des variables qualitatives avec plusieurs modalités. Pour le suivi *in situ*, il s'agissait :

- du type de prélèvement (2 classes : prélèvements ponctuels ou bilans 24h),
- des 33 dispositifs de traitement classés en 13 filières et 3 familles c'est-à-dire 33 classes, 13 classes ou 3 classes selon les échelles d'analyse retenues,
- de l'âge des installations au moment du prélèvement réparti en 3 classes :
 - < 2 ans,
 - entre 2 et 4 ans,
 - 4 ans et plus.
- de leur taux de charge au moment du prélèvement réparti en 3 classes :
 - < 30 %,
 - entre 30 % et 70 %,
 - > 70%.

Il est important de considérer des variables significatives indépendantes entre elles, cette obligation étant en lien avec la structuration même du modèle, sous une forme linéaire.

Pour identifier l'effet de chaque variable explicative, celle-ci doit être codée par des indicateurs binaires. Le codage pour un **prélèvement ponctuel** réalisé sur le rejet d'un dispositif de **la famille A**, âgé **de plus de 4 ans** et d'un **taux de charge < 30 %** est présenté à titre d'exemple au Tableau 1.

Tableau 1 : Exemple de codage des variables explicatives par des variables indicatrices

Variable explicative	Modalité	Codage
Prélèvement	Bilan 24h	0
	Ponctuel	1
Famille	A	1
	B	0
	C	0
Âge	< 2 ans	0
	2 ans – 4 ans	0
	> 4 ans	1
Charge	< 30%	1
	30% – 70%	0
	> 70%	0

Dans le modèle, un ensemble de référence est choisi et est ensuite comparé aux autres modalités. Dans le suivi *in situ*, l'ensemble de référence (Tableau 2) utilisé était composé de l'effectif le plus élevé observé parmi les modalités de chaque variable explicative, soit ici l'un des deux choix de prélèvement, l'une des trois familles, l'une des trois classes d'âge et l'une des trois classes de taux de charge. **Le choix de la référence peut être adapté à l'étude de l'utilisateur, et n'influe pas sur les conclusions du modèle.**

Ensuite, les effets des autres variables explicatives par rapport à cet ensemble de référence seront testés et donneront les coefficients correspondants ($\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \dots$), ainsi que les coefficients ($\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7, \dots$) pour les effets sur la dispersion.

Tableau 2 : Exemple d'un ensemble de référence et des variables explicatives testées

		Ensemble de référence	Variables explicatives	Coefficients associés
Prélèvement	Bilan 24h		X_1	β_1, γ_1
	Ponctuel	Oui		
Famille	A	Oui		
	B		X_2	β_2, γ_2
	C		X_3	β_3, γ_3
Age	< 2 ans		X_4	β_4, γ_4
	2 ans – 4 ans		X_5	β_5, γ_5
	> 4 ans	Oui		
Charge	< 30 %	Oui		
	30 % – 70 %		X_6	β_6, γ_6
	> 70 %		X_7	β_7, γ_7

Un test d'hypothèse de la significativité des effets des 7 variables explicatives est réalisé :

H_0 : « l'effet de la modalité considérée n'est pas significativement différent de celui de la référence »

Ce test calcule la probabilité p (appelée p -valeur) de rejeter à tort l'hypothèse nulle H_0 . Si l'hypothèse H_0 est retenue, alors l'effet est considéré comme « non significatif ». L'hypothèse H_0 n'est pas acceptée si la p -valeur est inférieure au seuil de significativité α ; l'effet est alors considéré comme significatif. **Par convention, ce seuil de significativité est communément fixé à 5 % (MBENGUE, 2010).** Dans le cadre de l'étude *in situ*, la valeur retenue pour α a été volontairement faible (0,1 %) afin de ne retenir que des effets très solides (PY, 1999).

On procède alors par itération en démarrant le processus en considérant le modèle complet, puis en éliminant à chaque itération l'effet le moins significatif (p -valeur la plus élevée) parmi les effets non significatifs (p -valeur supérieure à 0,1 %), et ce jusqu'à ne conserver que des effets significatifs.

Le tableau 3 ci-dessous illustre un exemple d'utilisation du modèle.

Tableau 3 : Exemple de tests permettant d'identifier les effets significatifs et non significatifs des variables explicatives, pour des p -valeurs exceptionnellement fixées à 0,1 %

Ensemble ¹ de référence du 1 ^{er} test	
Prélèvement	Bilan 24h
Famille	B
Age	2 ans – 4 ans
Charge	> 70%

¹ L'ensemble de référence retenu est constitué des sous-ensembles d'effectif maximum de chaque variable explicative.

1 ^{er} test	Variables explicatives			Résultats du test d'hypothèse	
	Prélèvement	Ponctuel	X ₁	p > 0,1 %	non significatif
	Famille	A	X ₂	p < 0,1 %	significatif
		C	X ₃	p > 0,1 %	non significatif
	Age	< 2 ans	X ₄	p > 0,1 %	non significatif
		> 4 ans	X ₅	p < 0,1 %	significatif
	Charge	< 30%	X ₆	p > 0,1 %	non significatif
		30% – 70%	X ₇	p < 0,1 %	significatif

Ensemble de référence du dernier test			
Prélèvement	Bilan 24h	et	ponctuel
Famille	B	et	C
Age	2 ans – 4 ans	et	< 2 ans
Charge	> 70%	et	< 30 %

Dernier test	Variables explicatives			Résultats du test d'hypothèse	
	Famille	A	X ₁	p < 0,1 %	significatif
	Age	> 4 ans	X ₂	p < 0,1 %	significatif
	Charge	30 % – 70 %	X ₃	p < 0,1 %	significatif

A partir de cet exemple, les conclusions sont les suivantes :

- Le mode de prélèvement n'a pas d'effet sur la variable dépendante Y, et l'outil considère que les deux types de prélèvement sont semblables.
- Pour les familles,
 - La famille A n'est pas semblable aux familles B et C,
 - L'outil n'identifie pas d'effet significatif distinguant les familles B et C.
- Pour les classes d'âge,
 - Les dispositifs mis récemment en service (moins de 4 ans) ne sont pas semblables aux dispositifs mis en service depuis plus de 4 ans,
 - L'outil n'identifie pas d'effet de l'âge du dispositif dès lors que celui-ci a moins de 4 ans.
- Pour le taux de charge,
 - Les dispositifs moyennement chargés (30 % – 70 %) ne sont pas semblables aux dispositifs faiblement (< 30 %) ou fortement (> 70 %) chargés,
 - L'outil n'identifie pas d'effet de charge dès lors que le dispositif est, soit faiblement (< 30 %), soit fortement (> 70 %) chargé.

Lorsque les effets des variables explicatives sont déterminés, le modèle permet d'obtenir les médianes de chaque ensemble par rapport à la variable dépendante :

$$Y = e^{\mu}$$

De plus, les coefficients β_1, β_2, \dots déterminent des facteurs quantifiant l'influence (positive ou négative) attachés à chaque variable explicative X_1, X_2, \dots par rapport à la référence. Ils se calculent ainsi :

- si β est positif, le facteur de modification est $e^{\beta} - 1$. Par exemple :
Pour $\beta = 1,5$, le facteur est $e^{1,5} - 1$, soit 3,48 ou une augmentation de 348 %.
- si β est négatif, le facteur de modification est $1 - e^{\beta}$. Par exemple :

Pour $\beta = -1,5$, le facteur est $1 - e^{-1,5}$, soit 0,77 ou une diminution de 77 %.

Cette démarche est alors appliquée à toutes les variables dépendantes (MES, DCO, DBO₅, NK, N-NH₄⁺, N-NO₃⁻).

➤ Utilisation pratique à partir d'un autre cas d'étude que le suivi *in situ*

Script de base à utiliser : le fichier R « GlnLM-EffLocDisp-DT-fctNM-01_2021 »

Ce script contient le modèle ln-linéaire généralisé prenant en compte les effets sur la dispersion. La gestion des données est réalisée en DataTable et la méthode d'optimisation de Nelder-Mead est utilisée pour le modèle.

Données d'entrée à fournir au script : un document au format csv (séparateur : point-virgule) contenant les variables dépendantes à analyser (ex : MES, DCO, ...) ainsi que les différentes variables explicatives (ex : Dispositif, Classe d'âge, ...) avec les titres correspondants en tête de colonne.

Exemple de fichier : BDD_exemple_lin-lineaire.csv

Dans cet exemple, on s'intéresse aux données correspondant aux rejets de 3 dispositifs d'assainissement des eaux usées différents. Les variables dépendantes analysées sont les concentrations de sortie en MES, DCO, DBO₅, NTK, N-NH₄⁺ et N-NO₃⁻. On cherche à étudier l'impact sur ces concentrations des variables explicatives suivantes :

- Le dispositif (D0, D1 ou D2) ;
- L'âge de l'installation au moment du prélèvement (< 2 ans ou > 2 ans) ;
- Le taux de charge de l'installation (< 70 % ou > 70 %).

Utilisation du script :

```
1 #####
2 ## Modèle ln-linéaire généralisé
3 ## Gestion données en DataTable - optimisation Nelder-Mead en fonction
4 ## 2020-01-20
5 ## source("ANC/GlnLM-EffLocDisp-DT-fctNM.R")
6 #####
7
8 options(width=100);
9 ok <- 1;
10 if (ok == 1) {
11   rm(list = ls());
12 }
13
14 library(data.table);
15 library(numDeriv);
16
17 ## Répertoire de travail ##
18 chem <- "/Users/eva.falipou/Documents/Eva/ANC/ln_linéaire/";
19
```

Facultatif : définir un chemin vers le répertoire de travail où se trouvent les données (ligne 18 du script). Cela permet d'éviter d'avoir à renseigner à nouveau le chemin dans le cas d'usage multiple de l'outil.

Pour adapter le code à d'autres cas d'étude, l'essentiel des modifications est à faire dans le bloc « Lecture données ». Selon les paramètres analysés, il faut penser à introduire la LQ correspondant aux conditions analytiques du paramètre.

Une autre modification du script consiste à éventuellement changer le quantile des résultats à afficher (voir la partie « interprétation des résultats »).

Le code ci-dessous est un exemple visant à étudier l'effet des variables explicatives citées plus haut sur les concentrations en NTK mesurées en sortie des dispositifs étudiés.

L'utilisateur doit tout d'abord déterminer l'ensemble de référence, choisi ici (par exemple) comme correspondant à l'effectif le plus élevé observé parmi les modalités de chaque variable explicative pour le paramètre NTK. Il s'agit pour cet exemple de D0 pour le dispositif, de la classe « > 2 ans » pour l'âge et de la classe « > 70% » pour la charge.

```
20 ▾ #####
21 ## Lecture donnees - Exemple : analyse NTK avec log=1
22 ▾ #####
23 ok <-1;
24 ▾ if (ok == 1) {
25   fin <- paste(chem,"BDD_exemple_1n-lineaire.csv",sep="");
26   don <- fread(fin,dec=",",sep=";");
27   don[,y:=NTK]; # choix de la variable dependante
28   ## Libelle variable dependante pour documenter affichages ##
29   labvdep <- "NTK";
30   ## Left-censoring ##
31   log <- 1.0; # LQ correspondant a la variable dependante
32   don[,lb:=y];
33   don[y<=log,lb:=0];
34   don[,ub:=y];
35   don[y<=log,ub:=log];
36   ## Variables explicatives ##
37   don[,D0:=ifelse(Dispositif=="D0",1,0)];
38   don[,D2:=ifelse(Dispositif=="D2",1,0)];
39   don[,inf2ans:=ifelse(classe_age=="< 2 ans",1,0)];
40   don[,inf70pct:=ifelse(classe_charge=="< 70%",1,0)];
41   ## Nettoyage ##
42   don <- don[,.(y,lb,ub,D0,D2,inf2ans,inf70pct)];
43   don <- na.omit(don);
44   ## Effects on location (median) ##
45   zm <- c("D0");
46   #zm <- c();## No effect on location ##
47   nbzm <- length(zm);
48   ## Effects on dispersion ##
49   zd <- c();
50   #zd <- c("D1","D2","inf2ans","inf70pct");## No effect on dispersion ##
51   nbzd <- length(zd);
52 ▾ }
```

Le bloc « lecture des données » ci-dessus crée une DataTable appelée « don ».

La première étape est de récupérer le fichier contenant les données à analyser, de le lire et de le stocker dans la DataTable « don » (lignes 25 et 26).

On crée ensuite la colonne y correspondant à la variable dépendante à étudier, ici la concentration en NTK (ligne 27). Les lignes 30 à 35 correspondent à la création des colonnes lb et ub représentant les bornes de l'intervalle censuré à gauche, intervalle compris entre 0 et la LQ correspondant à celle de la variable dépendante, en l'occurrence 1,0 mg/L pour NTK (à définir ligne 31).

C'est au niveau des lignes 36 à 40 dans l'exemple que l'on code en binaire les modalités des variables explicatives à étudier (hors ensemble de référence). Elles sont ensuite intégrées à la DataTable « don » en tant que nouvelles colonnes. La ligne 39 par exemple correspond à la création d'une colonne, nommée « inf2ans » par l'utilisateur, contenant la valeur « 1 » pour chaque ligne de la colonne « classe_age » du fichier csv contenant l'expression « < 2 ans ».

La ligne 42 « nettoie » la DataTable pour ne laisser que les colonnes y, lb, ub et celles des modalités des variables explicatives hors ensemble de référence (soit ici « D1 », « D2 », « inf2ans » et « inf70pct »).

La ligne 43 supprime de plus les lignes pour lesquelles on ne dispose pas de valeur pour la variable dépendante (et contenant donc un « NA »).

En résumé, cette partie du script crée la DataTable « don » contenant :

- La variable dépendante à étudier (ici NTK) en colonne y ;
- La borne inférieure de l'intervalle des données censurées en colonne lb ;
- La borne supérieure de l'intervalle des données censurées en colonne ub ;
- Et enfin les différentes modalités des variables explicatives à étudier, à l'exception de celles prises en référence (soit ici les dispositifs D1 et D2, la classe d'âge « < 2 ans » et la classe de taux de charge « < 70 % »). Ces modalités sont codées en binaire suivant le principe présenté dans le tableau 1.

Pour mieux visualiser le résultat, voici un extrait de la DataTable « don » passé ces étapes :

▲	y	lb	ub	D1	D2	inf2ans	inf70pct
1	16.80	16.80	16.80	1	0	1	1
2	21.50	21.50	21.50	1	0	1	1
3	35.10	35.10	35.10	1	0	0	1
4	38.90	38.90	38.90	0	1	0	1
5	2.00	2.00	2.00	0	0	0	1
6	2.50	2.50	2.50	0	0	1	1
7	3.00	3.00	3.00	0	0	0	1
8	5.20	5.20	5.20	0	1	0	1
9	6.40	6.40	6.40	0	0	0	1
10	6.80	6.80	6.80	0	1	0	1
11	9.60	9.60	9.60	0	1	1	1
12	14.07	14.07	14.07	0	0	0	1
13	15.80	15.80	15.80	0	0	1	1
14	17.70	17.70	17.70	0	0	1	1
15	21.70	21.70	21.70	0	0	0	1
16	30.40	30.40	30.40	0	1	1	1
17	36.40	36.40	36.40	0	1	0	1
18	40.70	40.70	40.70	0	1	0	1
19	45.60	45.60	45.60	0	1	0	1
20	70.80	70.80	70.80	0	1	0	1
21	1.00	0.00	1.00	0	0	0	1
22	1.00	0.00	1.00	0	0	0	1
23	1.66	1.66	1.66	0	0	0	1
24	2.70	2.70	2.70	0	0	1	1
25	5.60	5.60	5.60	0	1	0	1
26	10.00	10.00	10.00	0	1	0	1
27	15.00	15.00	15.00	0	1	1	1
28	15.40	15.40	15.40	0	1	1	1
29	2.30	2.30	2.30	0	1	0	1
30	7.60	7.60	7.60	0	1	0	1

L'outil crée également les vecteurs zm et zd contenant la liste des modalités des variables explicatives à prendre en compte pour l'établissement du modèle, en distinguant celles considérées pour la médiane dans le vecteur zm, et celles considérées pour la dispersion dans le vecteur zd (lignes 44 et 48).

Les vecteurs zm et zd contiennent au début toutes les modalités des variables explicatives à comparer à l'ensemble de référence. Au fur et à mesure des tests, les modalités avec des effets non significatifs sont progressivement enlevées de ces vecteurs pour qu'ils ne contiennent plus que celles présentant des effets significatifs sur la médiane ou la dispersion. Enlever une modalité de ces vecteurs revient à l'agréger à la référence.

1^{er} test :

Le code tel que présenté plus haut est lancé et fournit les résultats suivants :

Dependent variable : NTK						
Label	Estimate	Std. Dev.	Ref	Chi2	DF	Pval
beta0	1.6221e+00	2.2711e-01	0	5.1017e+01	1	0.0000
D1	1.9404e+00	3.0811e-01	0	3.9662e+01	1	0.0000
D2	1.3927e+00	3.1602e-01	0	1.9422e+01	1	0.0000
inf2ans	-2.0789e-01	2.6726e-01	0	6.0508e-01	1	0.4366
inf70pct	-1.1105e-01	2.6124e-01	0	1.8068e-01	1	0.6708
gamma0	3.8101e-01	1.4172e-01	0	7.2276e+00	1	0.0072
D1	-6.7107e-01	2.9328e-01	0	5.2357e+00	1	0.0221
D2	-2.0959e-01	2.1933e-01	0	9.1317e-01	1	0.3393

inf2ans	6.8346e-03	2.5838e-01	0	6.9968e-04	1	0.9789
inf70pct	-2.5642e-01	2.0406e-01	0	1.5790e+00	1	0.2089

Ce tableau fournit les coefficients estimés par le modèle dans la colonne « Estimate », soit premièrement les coefficients (β_1, β_2, \dots) correspondant aux effets des variables explicatives sur la médiane de la variable dépendante, puis les coefficients ($\gamma_1, \gamma_2, \dots$) correspondant aux effets des variables explicatives sur la dispersion :

- La ligne beta0 correspond à la référence pour la médiane. Les lignes en-dessous correspondent aux différentes modalités des variables explicatives testées sur la médiane par rapport à la référence.
- De la même manière, la ligne gamma0 correspond à la référence pour la dispersion et les lignes en-dessous aux modalités des variables explicatives testées sur la dispersion par rapport à la référence.

On regarde alors au niveau de la colonne des p-values pour identifier les effets significatifs. Dans le suivi *in situ*, le seuil de significativité α était fixé à 0,1 %. En conservant cette valeur, on obtient ici un effet significatif des deux modalités « D1 » et « D2 » sur la médiane (surlignées en jaune ci-dessus).

Il est conseillé ensuite de retirer les effets non significatifs un par un par ordre de p-value décroissante (il faudrait commencer ici par enlever « inf2ans » sur la dispersion), en les intégrant dans la référence (c'est-à-dire en pratique en les enlevant des vecteurs zm ou zd) et en relançant le code. **Il est important de ne pas modifier le tableau « don », sinon les modèles successifs ne seront plus comparables s'ils ne concernent plus les mêmes échantillons.** Il se peut que, pas à pas, tous les effets disparaissent, ne laissant que les intercepts beta0 et gamma0.

Les tests s'arrêtent quand il ne reste dans les vecteurs zm et zd que des variables avec des effets significatifs, ou quand ces deux vecteurs sont vides.

Dernier test :

Avec cet exemple, après avoir enlevé un à un les effets non significatifs, on obtient les résultats suivants :

Dependent variable : NTK						
Label	Estimate	Std. Dev.	Ref	Chi2	DF	Pval
beta0	1.6219e+00	1.6698e-01	0	9.4352e+01	1	0.0000
D1	1.6938e+00	3.2286e-01	0	2.7523e+01	1	0.0000
D2	1.3215e+00	2.9949e-01	0	1.9470e+01	1	0.0000
gamma0	1.3044e-01	8.2280e-02	0	2.5134e+00	1	0.1129

Il ne reste plus que les modalités « D1 » et « D2 » identifiées comme ayant un effet significatif sur la médiane pour le paramètre NTK (p-value < 0,1%, surlignées en jaune ci-dessus). L'outil conclue donc que le dispositif D0 n'est ni semblable au dispositif D1, ni semblable au dispositif D2.

Il est à noter que ce test effectue des comparaisons uniquement vis-à-vis de la référence D0 et ne compare pas les modalités D1 et D2 entre elles. Pour répondre à cette question, une possibilité est de modifier la référence, en remplaçant D0 par D1 (ou D2), et de relancer le modèle. On obtient alors les résultats ci-dessous :

Avant-dernier test :

Dependent variable : NTK						
Label	Estimate	Std. Dev.	Ref	Chi2	DF	Pval
beta0	3.3153e+00	2.7635e-01	0	1.4393e+02	1	0.0000
D0	-1.6939e+00	3.2288e-01	0	2.7521e+01	1	0.0000
D2	-3.7148e-01	3.7174e-01	0	9.9859e-01	1	0.3177
gamma0	1.3051e-01	8.2286e-02	0	2.5156e+00	1	0.1127

Dernier test :

Dependent variable : NTK						
Label	Estimate	Std. Dev.	Ref	Chi2	DF	Pval
beta0	3.1101e+00	1.8608e-01	0	2.7934e+02	1	0.0000
D0	-1.4892e+00	2.5079e-01	0	3.5259e+01	1	0.0000
gamma0	1.3723e-01	8.2274e-02	0	2.7823e+00	1	0.0953

On voit que la p-value n'est pas significative au niveau de la modalité D2. L'outil conclue donc que les dispositifs D1 et D2 ne sont pas différents. Le dispositif D0 est le seul à se distinguer.

Point de vigilance

Il est possible que le modèle ne converge pas et renvoie alors le message « maxiter reached », ou bien qu'il ne puisse pas calculer d'écart-type d'estimation et ne fournisse donc pas de p-value (renvoyant ainsi des NA dans le tableau de résultats). Cette situation peut être rencontrée dans le cas de jeux de données très réduits, et le modèle n'est alors pas applicable.

Interprétation des résultats : comparaison des médianes (ou d'un autre percentile) des distributions distinguées

Les médianes correspondant à la distribution théorique calculée par l'outil peuvent être retrouvées à partir des premières lignes de la colonne « estimate » (jusqu'à gamma0) du premier tableau de résultats renvoyé, qui correspondent aux coefficients β_1, β_2, \dots . Ce calcul se fait directement grâce à l'expression $Y = e^{\beta_0 + \beta_1.X_1 + \dots}$ (voir la première partie du document présentant le principe de la méthode).

On considère le résultat final de l'exemple développé plus haut :

Dependent variable : NTK						
Label	Estimate	Std. Dev.	Ref	Chi2	DF	Pval
beta0	3.1101e+00	1.8608e-01	0	2.7934e+02	1	0.0000
D0	-1.4892e+00	2.5079e-01	0	3.5259e+01	1	0.0000
gamma0	1.3723e-01	8.2274e-02	0	2.7823e+00	1	0.0953

Les conclusions de l'outil permettent ici de distinguer statistiquement le dispositif D0 des deux autres pour le paramètre NTK. La colonne « estimate » permet d'obtenir les médianes calculées par l'outil pour la distribution des données du dispositif D0, et pour la distribution de référence, c'est-à-dire des données des dispositifs D1 et D2 (non distingués statistiquement) :

Médiane NTK (D1+D2) = $e^{\beta_0} = e^{3,1101} = 22,4$ mg/L

Médiane NTK (D0) = $e^{\beta_0 + \beta_{D0}} = e^{3,1101 - 1,4892} = 5,1$ mg/L

Les coefficients β_1, β_2, \dots de la colonne « estimate » peuvent également permettre de déterminer des facteurs quantifiant l'influence (positive ou négative) attachés à chaque variable explicative X_1, X_2, \dots par rapport à la référence (voir « principe de la méthode »).

Dans l'exemple, le coefficient β lié à la modalité D0 étant négatif, le facteur de modification est :

$$1 - e^{\beta_{D0}} = 1 - e^{-1,4892} = 0,77$$

Soit une diminution de 77 % par rapport à la référence.

En plus du tableau de résultats permettant d'identifier les variables ayant un effet significatif, le script fournit en sortie un second tableau contenant les médianes (ou un autre percentile choisi par l'utilisateur) observées et calculées par l'outil, pour tous les sous-groupes de données construits à partir des variables ayant un effet identifié.

L'utilisateur peut choisir le percentile à afficher dans le dernier bloc du script, « quantile estimation ». Il s'agit de la variable nommée « percent » :

```

306 - #####
307 - ## Quantile estimation
308 - #####
309 - percent <- 0.5; ## Quantile probability ##
310 - ## Quantiles are estimated by fac variables ##
311 - fac <- unique(c(zm,zd));
312 - ## Empirical (observed) quantiles ##
313 - qua <- don[,.(qobs=quantile(ub,probs=percent,na.rm=T),nbobs=.N),by=fac];
314 - n <- qua[,.N];
315 - ## Regression - Effects on median ##
316 - if (nbzm > 0) {
317 -   Zm <- cbind(array(1,dim=c(n,1)),as.matrix(qua[,zm,with=F]));
318 - } else {
319 -   Zm <- array(1,dim=c(n,1));
320 - }
321 - zbm <- Zm%*%thopt[1:(1+nbzm)];
322 - ## Regression - Effects on dispersion ##
323 - if (nbzd > 0) {
324 -   Zd <- cbind(array(1,dim=c(n,1)),as.matrix(qua[,zd,with=F]));
325 - } else {
326 -   Zd <- array(1,dim=c(n,1));
327 - }
328 - zbd <- Zd%*%thopt[(2+nbzm):(2+nbzm+nbzd)];

```

Dans le cas ci-dessus, « percent » est fixé à 0,5 soit la médiane (50^{ème} percentile). Fixer « percent » à 0,8, par exemple, permettrait de générer les 80^{èmes} percentiles.

Pour le même exemple, on obtient les résultats suivants :

Dependent variable : NTK - Table of observed and theoretical 50%-quantiles:

	D0	nbobs	nbfce	qobs	qcal	vqc	lbqc	ubqc
1:	0	38	0	29.75	22.42336	0.03462717	15.570515	32.292256
2:	1	48	8	6.10	5.05783	0.02826795	3.637881	7.032018

Les premières colonnes correspondent aux variables significatives restantes à la fin des tests, codées en binaire (ici uniquement D0). Les colonnes nbobs et nbfce correspondent respectivement au nombre de données total et au nombre de données censurées à gauche dans chaque sous-ensemble. La colonne qobs correspond au percentile réellement observé sur le jeu de données, et la colonne qcal au percentile de la distribution théorique déterminée par l'outil. Ces deux résultats sont fournis pour les différents sous-ensembles (2 ici) définis en fonction des variables significatives restantes. Ils sont reportés dans le tableau suivant pour plus de lisibilité :

	Médiane observée	Médiane calculée
Référence (D1+D2)	29,75	22,42
D0	6,10	5,06

On retrouve bien les résultats calculés manuellement à partir des « estimate ». L'intérêt est de pouvoir demander à l'outil de fournir un percentile différent de la médiane.

Enfin, les trois dernières colonnes fournissent des informations statistiques correspondant respectivement à la variance et aux bornes d'estimation inférieures et supérieures du percentile théorique.

Des différences peuvent exister entre les percentiles observés et calculés, liées à la qualité de l'estimation réalisée par le modèle. Ces différences sont susceptibles d'apparaître et peuvent s'expliquer par des différences de dispersion entre les sous-ensembles ou par les effectifs (trop restreints) de certains sous-ensembles.

Il est nécessaire dans cette situation-là de réaliser un tableau de contingence pour identifier les effectifs de chaque catégorie. Par exemple, dans un cas où l'outil identifie un effet significatif conjoint du type de dispositif (en distinguant D0 par rapport aux deux autres), de l'âge et du taux de charge sur le paramètre MES :

Nombre de données	Taux de charge < 70%		Taux de charge > 70%	
	Âge < 2 ans	Âge > 2 ans	Âge < 2 ans	Âge > 2 ans
D0	15	6	18	3
(D1 + D2)	34	20	27	23

On voit que deux catégories sont représentées par un nombre très faible de données (en rouge). La notion de taille minimale en statistiques est complexe et 30 unités est la plus petite taille d'échantillon répondant à la loi des grands nombres. Cette règle empirique est retenue unanimement par la communauté des statisticiens pour construire les plans d'expériences. Pour des échantillons d'effectif restreint, il convient d'exercer un regard critique mais l'analyse par l'outil est possible.

A titre informatif, l'analyse réalisée lors du suivi *in situ* avait permis de retenir une taille de 13 valeurs minimum, cette valeur étant déterminée par essais-erreurs.

Autre exemple avec des effets conjoints, c'est à dire plusieurs effets simultanés identifiés

L'exemple développé plus haut étant relativement simple, un nouveau cas est détaillé ici pour assurer une bonne compréhension de l'analyse des résultats.

Les variables explicatives sont les mêmes mais l'outil est utilisé sur un nouveau jeu de données.

Voici le résultat final du test réalisé sur le paramètre MES :

Dependent variable : MES						
Label	Estimate	Std. Dev.	Ref	Chi2	DF	Pval
beta0	2.2062e+00	1.0535e-01	0	4.3854e+02	1	0.0000
D1	-6.4239e-01	1.2062e-01	0	2.8364e+01	1	0.0000
sup70pct	7.3304e-01	1.2595e-01	0	3.3871e+01	1	0.0000
gamma0	9.5772e-02	6.7283e-02	0	2.0261e+00	1	0.1546
D1	-6.0292e-01	1.0864e-01	0	3.0797e+01	1	0.0000

L'outil identifie un effet conjoint du dispositif et du taux de charge sur la médiane, ainsi qu'un effet du dispositif sur la dispersion (avec une moindre dispersion obtenue avec le dispositif 1).

Les médianes calculées par l'outil peuvent être retrouvées à partir des premières lignes de la colonne « estimate » (jusqu'à gamma0) de ce tableau de résultats, qui correspondent aux coefficients β_1 , β_2 , ...

Par exemple, la médiane calculée pour les données correspondant au dispositif D1 pour des taux de charge supérieurs à 70 % est : $e^{\beta_0 + \beta_{D1} + \beta_{sup70pct}} = e^{(2,2062-0,64239+0,73304)} = 9,94$

Ainsi, on obtient le tableau suivant :

Médiane calculée		Taux de charge	
		< 70 % (ref)	> 70 %
Dispositif	D1	$e^{\beta_0 + \beta_{D1}} = e^{(2,2062-0,64239)} = 4,78$	$e^{\beta_0 + \beta_{D1} + \beta_{sup70pct}} = e^{(2,2062-0,64239+0,73304)} = 9,94$
	D0+D2 (ref)	$e^{\beta_0} = e^{(2,2062)} = 9,08$	$e^{\beta_0 + \beta_{sup70pct}} = e^{(2,2062+0,73304)} = 18,9$

Le script fournit de plus le tableau suivant contenant les différentes médianes (pour percent=0,5) :

Dependent variable : MES - Table of observed and theoretical 50%-quantiles:										
	D1	sup70pct	nbobs	nb1ce	qobs	qcal	vqc	lbqc	ubqc	
1:	0	0	90	5	10	9.081419	0.011099322	7.387120	11.164320	
2:	0	1	33	3	24	18.902071	0.018417737	14.487353	24.662083	
3:	1	0	58	6	5	4.777123	0.005809343	4.114219	5.546836	
4:	1	1	22	1	10	9.943106	0.012912448	7.957850	12.423627	

Les médianes théoriques et calculées par l'outil sont fournies pour les différents sous-ensembles (4 ici) définis en fonction des variables significatives restantes. On obtient donc ici :

Médiane observée (médiane calculée)		Taux de charge	
		< 70 % (ref)	> 70 %
Dispositif	D1	5 (4,777123)	10 (9,943106)
	D0+D2 (ref)	10 (9,081419)	24 (18,902071)

En conclusion, l'outil identifie un effet du dispositif et du taux de charge sur le paramètre MES : les concentrations en MES sont moins élevées avec le dispositif D1, et plus élevées avec un taux de charge supérieur à 70 %.

➤ Utilisation de la version développée sous R-Shiny

Le script du modèle présenté plus haut a également été adapté sous la forme d'une « Shiny App », c'est-à-dire une application web interactive.

Elle est accessible depuis le dépôt gitlab de l'unité REVERSAAL à l'adresse :

<https://gitlab.irstea.fr/reversaal/loglinear-regression>

Pour la faire fonctionner l'utilisateur doit récupérer le script « generalized ln-linear.R » et le document R markdown « results.Rmd » nécessaire à l'exportation des résultats.

L'application se lance par l'intermédiaire du bouton « Run App » dans le script :



L'application présente une interface plus agréable à utiliser que le script brut bien que le principe reste le même. Le premier onglet « tutorial » présente la marche à suivre pour la faire fonctionner.

Generalized In-linear

Create data table

Choose csv file with semicolon separator

Browse...

No file selected

Quantification limit of the selected parameter

1

Number of explanatory variable modalities

2

	Label	Column header	Expression
mod 1			
mod 2			

Create

Effects on median

Effects on dispersion

Choose quantile (%)

Quantile

50

Show results

Download results

Document title

Results

Download

Tutorial

Data Table

Results

How to use the Generalized In-linear Shiny App

> Select the csv file containing the data to be analyzed

It has to be a csv file with semicolon separator

The file should contain the different dependent variables (ex : TSS, CDO, BDO,...) and the different explanatory variables with corresponding column headers (ex: facility, age group,...)

> Choose the dependent variable

For example TSS

> State the value of the corresponding quantification limit

For example 2.0 (mg/L) for TSS

> Indicate the number of explanatory variable modalities to be investigated

Other than those corresponding to the reference set

> Complete the table describing each explanatory variable modality with the following information:

- the label of the modality that you want to see in the different results tables (for example 'inf2ans')
- the column header in your csv file corresponding to the explanatory variable (for example 'classe_age')
- the expression of the modality inside this column in your csv file (for example '< 2 ans')

> Click on the 'Create' button and then check in the 'Data Table' panel if the table is well-built

> Choose the modalities to consider in order to evaluate their effects on median and dispersion

> Choose the quantile you want to be returned by the model

> Launch the model by clicking on the 'Show results' button. The table of estimates and the table of quantiles appear in the 'Results' panel.

> Export results

It is finally possible to comment the results (dataset used, reference set, conclusions,...) and to export them in a word file using the 'Download' button (with a title chosen by the user).

Un exemple d'utilisation est développé ci-dessous. Il s'agit du même que celui développé dans la partie « utilisation du modèle ».

Les premières étapes permettent de créer la DataTable utilisée dans le modèle en :

- téléchargeant le fichier csv contenant les données,
- sélectionnant le paramètre à étudier parmi les colonnes du fichier,
- renseignant la limite de quantification liée à ce paramètre,
- et enfin en complétant un tableau décrivant les modalités des variables explicatives (hors ensemble de référence) à étudier.

La capture d'écran ci-dessous montre les différents champs complétés pour correspondre à l'exemple développé, soit l'étude de l'impact sur la concentration en NTK (limite de quantification = 1,0 mg/L) des variables explicatives suivantes :

- Le dispositif (D0, D1 ou D2) ;
- L'âge de l'installation au moment du prélèvement (< 2 ans ou > 2 ans) ;
- Le taux de charge de l'installation (< 70 % ou > 70 %).

L'ensemble de référence correspond aux modalités : D0 pour le dispositif, « > 2 ans » pour l'âge et « > 70% » pour la charge.

Une fois les champs remplis, le bouton « Create » permet de créer la DataTable.

Generalized In-linear

Create data table

Choose csv file with semicolon separator

Browse...

BDD_exemple_In-lineaire.csv

Upload complete

Select parameter

NTK

Quantification limit of the selected parameter

1

Number of explanatory variable modalities

4

	Label	Column header	Expression
mod 1	D1	Dispositif	D1
mod 2	D2	Dispositif	D2
mod 3	inf2ans	classe_age	< 2 ans
mod 4	inf70pct	classe_charge	< 70%

Create

Effects on median

Effects on median

☒ D1

☒ D2

☒ inf2ans

☒ inf70pct

Effects on dispersion

Effects on dispersion

☒ D1

☒ D2

☒ inf2ans

☒ inf70pct

Choose quantile (%)

Quantile

50

Show results

Tutorial

Data Table

Results

How to use the Generalized In-linear Shiny App

> Select the csv file containing the data to be analyzed

It has to be a csv file with semicolon separator

The file should contain the different dependent variables (ex : TSS, CDO, BDO,...) and the different explanatory variables with corresponding column headers (ex: facility, age group,...)

> Choose the dependent variable

For example TSS

> State the value of the corresponding quantification limit

For example 2.0 (mg/L) for TSS

> Indicate the number of explanatory variable modalities to be investigated

Other than those corresponding to the reference set

> Complete the table describing each explanatory variable modality with the following information:

- the label of the modality that you want to see in the different results tables (for example 'inf2ans')

- the column header in your csv file corresponding to the explanatory variable (for example 'classe_age')

- the expression of the modality inside this column in your csv file (for example '< 2 ans')

> Click on the 'Create' button and then check in the 'Data Table' panel if the table is well-built

> Choose the modalities to consider in order to evaluate their effets on median and dispersion

> Choose the quantile you want to be returned by the model

> Launch the model by clicking on the 'Show results' button. The table of estimates and the table of quantiles appear in the 'Results' panel.

> Export results

It is finally possible to comment the results (dataset used, reference set, conclusions,...) and to export them in a word file using the 'Download' button (with a title choosen by the user).

L'utilisateur peut contrôler la DataTable créée dans l'onglet « Data Table » comme on le voit ci-dessous :

Generalized In-linear

Create data table

Choose csv file with semicolon separator

Select parameter

Quantification limit of the selected parameter

Number of explanatory variable modalities

	Label	Column header	Expression
mod 1	D1	Dispositif	D1
mod 2	D2	Dispositif	D2
mod 3	inf2ans	classe_age	< 2 ans
mod 4	inf70pct	classe_charge	< 70%

Effects on median

Effects on median

☒ D1

☒ D2

☒ inf2ans

☒ inf70pct

Effects on dispersion

Effects on dispersion

☒ D1

☒ D2

☒ inf2ans

☒ inf70pct

Choose quantile (%)

Quantile

[Tutorial](#)
[Data Table](#)
[Results](#)

Parameter: NTK

y	lb	ub	D1	D2	inf2ans	inf70pct
16.80	16.80	16.80	1.00	0.00	1.00	1.00
21.50	21.50	21.50	1.00	0.00	1.00	1.00
35.10	35.10	35.10	1.00	0.00	0.00	1.00
38.90	38.90	38.90	0.00	1.00	0.00	1.00
2.00	2.00	2.00	0.00	0.00	0.00	1.00
2.50	2.50	2.50	0.00	0.00	1.00	1.00
3.00	3.00	3.00	0.00	0.00	0.00	1.00
5.20	5.20	5.20	0.00	1.00	0.00	1.00
6.40	6.40	6.40	0.00	0.00	0.00	1.00
6.80	6.80	6.80	0.00	1.00	0.00	1.00
9.60	9.60	9.60	0.00	1.00	1.00	1.00
14.07	14.07	14.07	0.00	0.00	0.00	1.00
15.80	15.80	15.80	0.00	0.00	1.00	1.00
17.70	17.70	17.70	0.00	0.00	1.00	1.00
21.70	21.70	21.70	0.00	0.00	0.00	1.00
30.40	30.40	30.40	0.00	1.00	1.00	1.00
36.40	36.40	36.40	0.00	1.00	0.00	1.00
40.70	40.70	40.70	0.00	1.00	0.00	1.00
45.60	45.60	45.60	0.00	1.00	0.00	1.00
70.80	70.80	70.80	0.00	1.00	0.00	1.00
1.00	0.00	1.00	0.00	0.00	0.00	1.00
1.00	0.00	1.00	0.00	0.00	0.00	1.00
1.66	1.66	1.66	0.00	0.00	0.00	1.00
2.70	2.70	2.70	0.00	0.00	1.00	1.00
5.60	5.60	5.60	0.00	1.00	0.00	1.00
10.00	10.00	10.00	0.00	1.00	0.00	1.00
15.00	15.00	15.00	0.00	1.00	1.00	1.00
15.40	15.40	15.40	0.00	1.00	1.00	1.00
2.30	2.30	2.30	0.00	1.00	0.00	1.00
7.60	7.60	7.60	0.00	1.00	0.00	1.00
21.20	21.20	21.20	1.00	0.00	1.00	1.00
32.00	32.00	32.00	1.00	0.00	0.00	1.00
34.90	34.90	34.90	1.00	0.00	1.00	1.00
42.00	42.00	42.00	1.00	0.00	0.00	1.00
46.00	46.00	46.00	1.00	0.00	1.00	1.00

Enfin, il est possible de sélectionner les modalités à prendre en compte dans le modèle au niveau du panneau de gauche : « Effects on median » et « Effects on dispersion ».

Une fois les modalités sélectionnées, l'utilisateur doit cliquer sur le bouton « Show result » pour lancer le programme et les résultats apparaitront dans l'onglet « Results ».

Le quantile renvoyé par l'outil pour les distributions observées et calculées est également modifiable (il est par défaut fixé à la médiane).

Les résultats du premier lancement du modèle sont présentés ci-dessous :

Create data table

Choose csv file with semicolon separator

Select parameter

Quantification limit of the selected parameter

Number of explanatory variable modalities

	Label	Column header	Expression
mod 1	D1	Dispositif	D1
mod 2	D2	Dispositif	D2
mod 3	inf2ans	classe_age	< 2 ans
mod 4	inf70pct	classe_charge	< 70%

Effects on median

Effects on median

☒ D1
☒ D2
☒ inf2ans
☒ inf70pct

Effects on dispersion

Effects on dispersion

☒ D1
☒ D2
☒ inf2ans
☒ inf70pct

Choose quantile (%)

Quantile

Download results

Document title

[Tutorial](#)
[Data Table](#)
[Results](#)

> Table of estimates

Parameter: NTK

Label	Estimate	Std_Dev	Ref	Chi2	DF	Pval
beta0	1.6221e+00	2.2711e-01	0	5.1017e+01	1	0.0000
D1	1.9404e+00	3.0811e-01	0	3.9662e+01	1	0.0000
D2	1.3927e+00	3.1602e-01	0	1.9422e+01	1	0.0000
inf2ans	-2.0789e-01	2.6726e-01	0	6.0508e-01	1	0.4366
inf70pct	-1.1105e-01	2.6124e-01	0	1.8068e-01	1	0.6708
gamma0	3.8101e-01	1.4172e-01	0	7.2276e+00	1	0.0072
D1	-6.7107e-01	2.9328e-01	0	5.2357e+00	1	0.0221
D2	-2.0959e-01	2.1933e-01	0	9.1317e-01	1	0.3393
inf2ans	6.8346e-03	2.5838e-01	0	6.9968e-04	1	0.9789
inf70pct	-2.5642e-01	2.0406e-01	0	1.5790e+00	1	0.2089

> Quantiles

Dependent variable: NTK - Table of observed and theoretical 50 % quantiles:

D1	D2	inf2ans	inf70pct	nbobs	nblice	qobs	qcal	vqc	lbqc	ubqc
0	0	0	0	34	6	7.69	5.064	0.0516	3.2446	7.9032
0	0	0	1	8	2	2.5	4.532	0.0706	2.6918	7.6291
0	0	1	0	1	0	33.8	4.113	0.1112	2.1398	7.9073
0	0	1	1	5	0	2.7	3.681	0.1099	1.9221	7.0498
0	1	0	0	4	0	50	20.386	0.096	11.1073	37.4157
0	1	0	1	11	0	10	18.243	0.0571	11.4207	29.1418
0	1	1	0	2	0	23	16.559	0.1419	7.9132	34.6527
0	1	1	1	4	0	15.2	14.819	0.0827	8.4329	26.0413
1	0	0	0	3	0	29.1	35.254	0.0785	20.3554	61.0559
1	0	0	1	3	0	35.1	31.548	0.0714	18.6886	53.2572
1	0	1	0	5	0	12	28.636	0.069	17.1122	47.9214
1	0	1	1	6	0	28.2	25.627	0.0415	17.1871	38.2103

Comment

On procède ensuite par itération en retirant l'effet le moins significatif (p-valeur la plus élevée) de la sélection des « Effects on median » ou « Effects on dispersion » et en relançant le modèle avec le bouton « Show results » jusqu'à ne conserver que des effets significatifs (p-valeur inférieure à 0,1 %) comme dans la capture d'écran ci-dessous.

Il est enfin possible d'ajouter un commentaire décrivant les résultats obtenus (ex : jeu de données utilisé, variables explicatives étudiées, ensemble de référence choisi, ...) et de télécharger les résultats dans un document Word qui contiendra le nom du paramètre étudié, le commentaire de l'utilisateur, le tableau des estimates et le tableau des quantiles (bouton « Download »).

Choose for the main estimation parameter

Browse... BDD_exemple_In-lineaire.csv

Upload complete

Select parameter

NTK

Quantification limit of the selected parameter

1

Number of explanatory variable modalities

4

	Label	Column header	Expression
mod 1	D1	Dispositif	D1
mod 2	D2	Dispositif	D2
mod 3	inf2ans	classe_age	< 2 ans
mod 4	inf70pct	classe_charge	< 70%

Create

Effects on median

Effects on median

☒ D1

☒ D2

☐ inf2ans

☐ inf70pct

Effects on dispersion

Effects on dispersion

☐ D1

☐ D2

☐ inf2ans

☐ inf70pct

Choose quantile (%)

Quantile

50

Show results

Download results

Document title

Exemple

Download

Parameter: NTK

Label	Estimate	Std_Dev	Ref	Chi2	DF	Pval
beta0	1.6219e+00	1.6698e-01	0	9.4352e+01	1	0.0000
D1	1.6938e+00	3.2286e-01	0	2.7523e+01	1	0.0000
D2	1.3215e+00	2.9949e-01	0	1.9470e+01	1	0.0000
gamma0	1.3044e-01	8.2280e-02	0	2.5134e+00	1	0.1129

> Quantiles

Dependent variable: NTK - Table of observed and theoretical 50 % quantiles:

D1	D2	nbobs	nbice	qobs	qcal	vqc	lbqc	ubqc
0	0	48	8	6.1	5.063	0.0279	3.6498	7.0231
0	1	21	0	30.4	18.981	0.0618	11.6595	30.8989
1	0	17	0	29.1	27.543	0.0764	16.025	47.3401

Comment

Jeu de données : "BDD_exemple_In-lineaire.csv", base utilisée dans l'exemple du guide d'utilisation du modèle In-linéaire

Variables explicatives : dispositif, âge, taux de charge

Ensemble de référence : D0, > 2 ans, > 70%