Variations géographiques et sociodémographiques de la liaison en français

Une modélisation statistique

Basilio Calderone
basilio.calderone@univ-tlse2.fr

CNRS
&
CLLE-ERSS Université de Toulouse-Jean-Jaurès

0. Plan

- Objectif du travail
- La liaison - une petite présentation
- Un corpus phonologique : PFC
- Les outils statistiques (RL et GAM)
- Le model final et le résultats
- Limites du travail et perspectives futures

0. Plan

- Objectif du travail
- La liaison - une petite présentation
- Un corpus phonologique : PFC
- Les outils statistiques (RL et GAM)
- Le model final et le résultats
- Limites du travail et perspectives futures

0.1 Préambule

1- Le present travail ne fait pas appelle aux théories/approches phonologiques sur la liaison en français ni sur des approches basées sur des contraintes syntaxiques (cadre HPSG et similaires) ou sur les aspects prosodiques

1- Le present travail ne fait pas appelle aux théories/approches phonologiques sur la liaison en français ni sur des approches basées sur des contraintes syntaxiques (cadre HPSG et similaires) ou sur les aspects prosodiques

2- Il est basé sur une analyse statistique des probabilités de réalisation (ou pas) des liaison dans un corpus d'observations qui espère être représentatif

1- Le present travail ne fait pas appelle aux théories/approches phonologiques sur la liaison en français ni sur des approches basées sur des contraintes syntaxiques (cadre HPSG et similaires) ou sur les aspects prosodiques

2- Il est basé sur une analyse statistique des probabilités de réalisation (ou pas) des liaison dans un corpus d'observations qui espère être représentatif

3-Il ne s'agit pas d'un modèle classifieur où les variables explicatives rentrent dans une boîte noire sans pouvoir les controller

1. Objectif

- Le present travail vise à réaliser un modèle prédictif de la liaison en français à partir d'un ensemble de variables sociolinguistiques, grammaticales et de surface

1. Objectif

- Le present travail vise à réaliser un modèle prédictif de la liaison en français à partir d'un ensemble de variables sociolinguistiques, grammaticales et de surface

- Dans notre approche on essaie de se baser sur les données observées pour synthétiser dans un modèle unique les variables explicatives qui sont significatives et participent à la réalisation (ou à la non réalisation) de la liaison

1. Objectif

- Le present travail vise à réaliser un modèle prédictif de la liaison en français à partir d'un ensemble de variables sociolinguistiques, grammaticales et de surface

- Dans notre approche on essaie de se baser sur les données observées pour synthétiser dans un modèle unique les variables explicatives qui sont significatives et participent à la réalisation (ou à la non réalisation) de la liaison

- Nous faisons des inférences statistiques : induire les caractéristiques inconnues d'une population (toutes les réalisations possibles de liaison) à partir d'un échantillon issu de cette population (notre jeu de données)

1. Objectif

- Le present travail vise à réaliser un modèle prédictif de la liaison en français à partir d'un ensemble de variables sociolinguistiques, grammaticales et de surface

- Dans notre approche on essaie de se baser sur les données observées pour synthétiser dans un modèle unique les variables explicatives qui sont significatives et participent à la réalisation (ou à la non réalisation) de la liaison

- Nous faisons des inférences statistiques : induire les caractéristiques inconnues d'une population (toutes les réalisations possibles de liaison) à partir d'un échantillon issu de cette population (notre jeu de données)

- Ce modèle nous permet d'identifier les variables qui ont un poids plus important parmi les variables prises en examen (choisies par moi même) et de les ordonner selon une échelle de "puissance explicative"

1. Objectif

- Le present travail vise à réaliser un modèle prédictif de la liaison en français à partir d'un ensemble de variables sociolinguistiques, grammaticales et de surface

- Dans notre approche on essaie de se baser sur les données observées pour synthétiser dans un modèle unique les variables explicatives qui sont significatives et participent à la réalisation (ou à la non réalisation) de la liaison

- Nous faisons des inférences statistiques : induire les caractéristiques inconnues d'une population (toutes les réalisations possibles de liaison) à partir d'un échantillon issu de cette population (notre jeu de données)

- Ce modèle nous permet d'identifier les variables qui ont un poids plus important parmi les variables prises en examen (choisies par moi même) et de les ordonner selon une échelle de "puissance explicative"

- Il s'agit d'un modèle prédictif, permettant de tirer des généralisations sur des données inconnues éventuelles

- Objectif du travail
- La liaison - une petite présentation
- Un corpus phonologique : PFC
- Les outils statistiques (RL et GAM)
- Le model final et le résultats
- Limites du travail et perspectives futures

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est \i.l‿ɛ\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est	\i.l‿ɛ\
nous avons	\nu.z‿a.vɔ̃\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est	\i.l‿ɛ\
nous avons	\nu.z‿a.vɔ̃\
les oiseaux	\le.z‿wa.zo\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est	\i.l‿ɛ\
nous avons	\nu.z‿a.vɔ̃\
les oiseaux	\le.z‿wa.zo\
petit avion	\pe.ti.t‿a.vjɔ̃\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est	\i.l‿ɛ\
nous avons	\nu.z‿a.vɔ̃\
les oiseaux	\le.z‿wa.zo\
petit avion	\pe.ti.t‿a.vjɔ̃\
trop avancé	\tʁo.p‿a.vɑ̃.se\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est	\i.l‿ɛ\
nous avons	\nu.z‿a.vɔ̃\
les oiseaux	\le.z‿wa.zo\
petit avion	\pe.ti.t‿a.vjɔ̃\
trop avancé	\tʁo.p‿a.vɑ̃.se\
pays imaginaires	\pe.i.z‿i.ma.ʒi.nɛʁ\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est	\i.l‿ɛ\
nous avons	\nu.z‿a.vɔ̃\
les oiseaux	\le.z‿wa.zo\
petit avion	\pe.ti.t‿a.vjɔ̃\
trop avancé	\tʁo.p‿a.vɑ̃.se\
pays imaginaires	\pe.i.z‿i.ma.ʒi.nɛʁ\
beaucoup à dire	\bo.ku.p‿a diʁ\

La liaison est un type de phénomène de *sandhi* qui consiste à prononcer la consonne finale d'un mot lorsque celui-ci précède un mot comportant une voyelle initiale.

Par exemple, entre le déterminant <les> /le/ et le substantif <amis> /ami/, tout locuteur natif insérera un phonème /z/ dit de liaison ⟶ \lɛz‿a.mi\

il est	\i.l‿ɛ\
nous avons	\nu.z‿a.vɔ̃\
les oiseaux	\le.z‿wa.zo\
petit avion	\pe.ti.t‿a.vjɔ̃\
trop avancé	\tʁo.p‿a.vɑ̃.se\
pays imaginaires	\pe.i.z‿i.ma.ʒi.nɛʁ\
beaucoup à dire	\bo.ku.p‿a diʁ\
pendant un	\pɑ̃.dɑ̃.t‿œ̃\

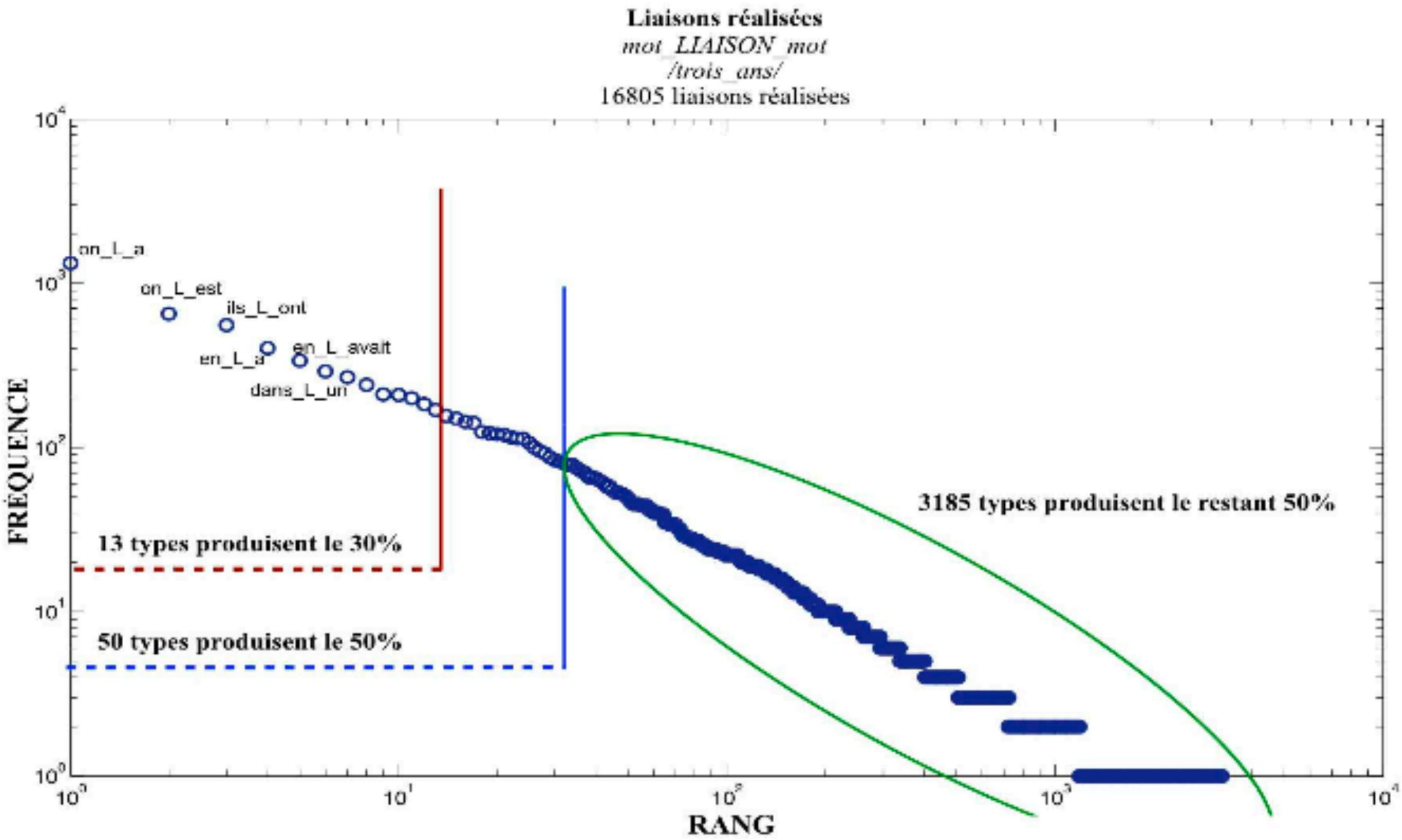
En littérature, on distingue a) la liaison obligatoire, b) la liaison facultative et c) la liaison interdite.

Liaison interdite **Liaison facultative** **Liaison obligatoire**

et encore pas encore les enfants

Analyse fréquentielle de la liaison en termes de *type* et *token* entre les deux mots (G et D)

De Jong 1996; Laks et al. 2015 et 2016



Analyse des correspondances (Benzécri, 1982) entre les POS des deux mots (G et D)
Laks et al. 2014

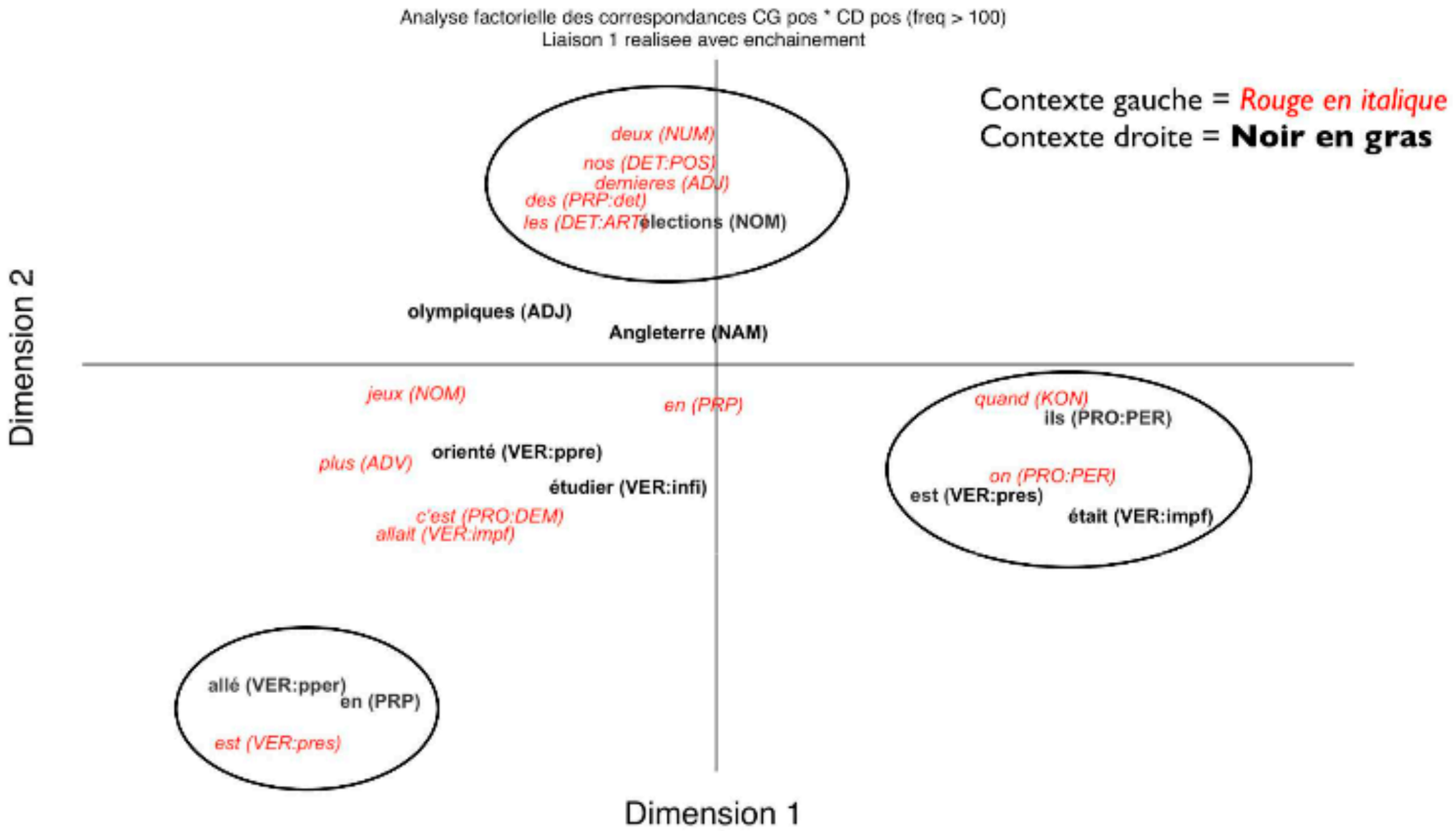
		POS DROITE									
			ADJ	**ADV**	**DET : ART**	**DET : POS**	**KON**	**NOM**	**INT**	**NUM**	**…**	**VER : ppres**
POS GAUCHE	**ADJ**	29	59	645	0	111	22	114	22	…	1
	ADV	444	7	12	98	31	76	32	201	…	99
	DET : ART	730	34	21	16	211	2	90	4	…	24
	DET : POS	36	87	12	23	1321	44	333	38	…	65
	KON	16	34	5	98	10	32	11	201	…	176
	NOM	1544	13	12	91	110	78	12	81	…	92
	INT	0	2	120	12	8	11	13	11	…	88
	NUM	24	80	70	5	3	55	32	20	…	67
	…	…	…	…	…	…	…	…	…	…	…
	VER : ppres	0	1	12	76	21	98	9	14	…	3

Analyse des correspondances (Benzécri, 1982) de la liaison en termes de *type* et *token* entre les deux mots (G et D)

Laks et al. 2014

Analyse des correspondances (Benzécri, 1982) de la liaison en termes de *type* et *token* entre les deux mots (G et D)

Laks et al. 2014

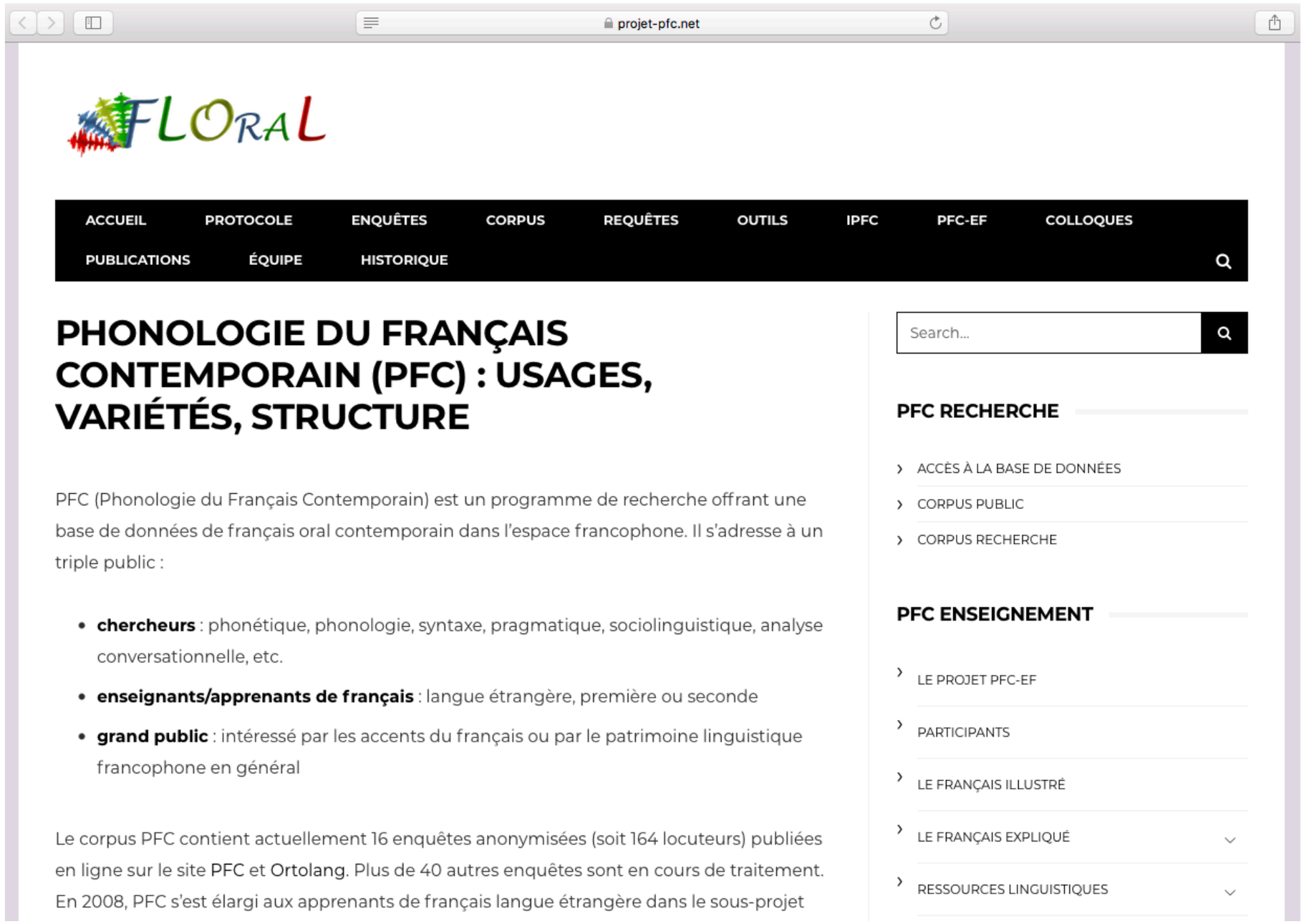


Dimension 1 + Dimension 2 = 47% variance

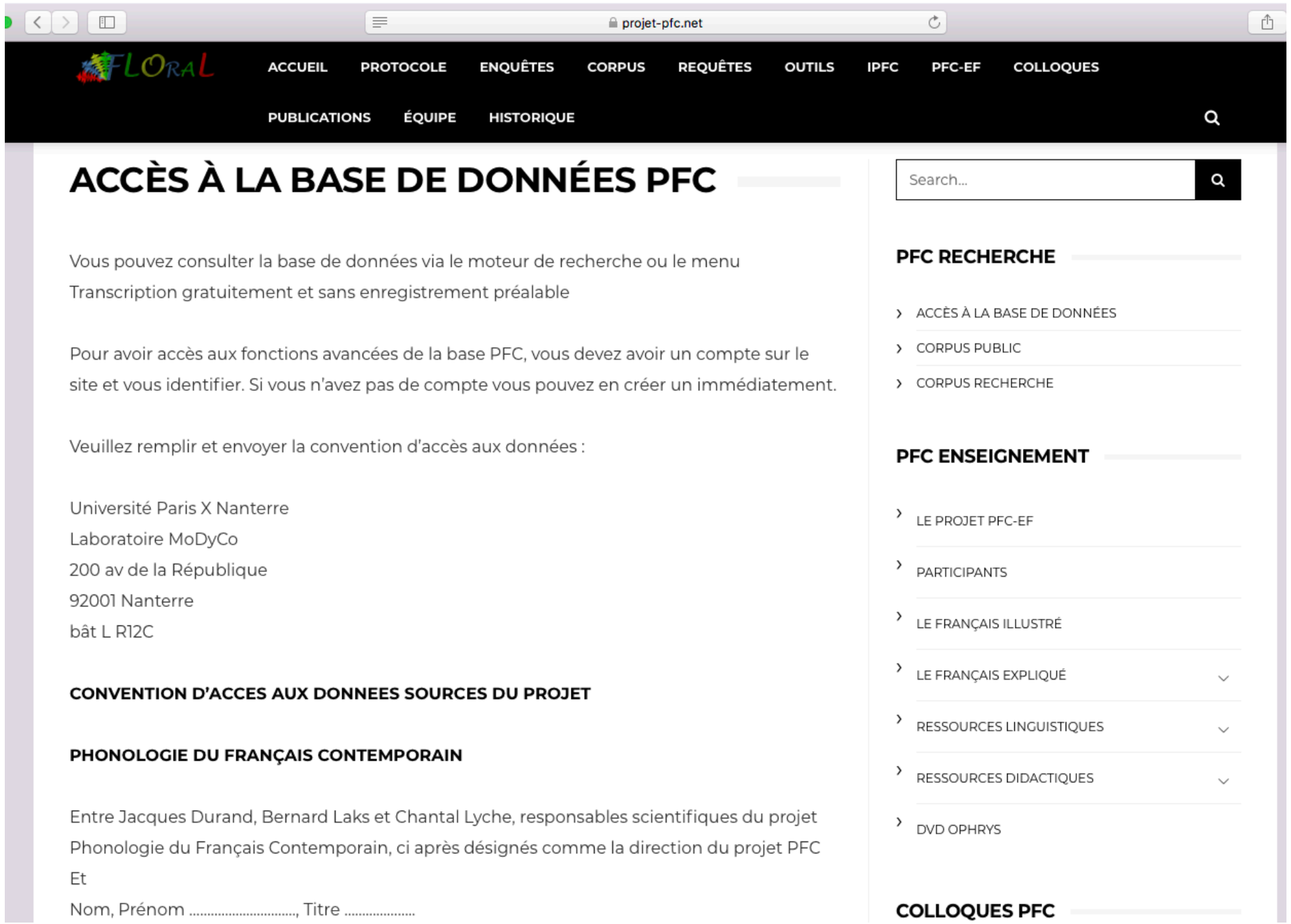
- Objectif du travail
- La liaison - une petite présentation
- Un corpus phonologique : PFC
- Les outils statistiques (RL et GAM)
- Le model final et le résultats
- Limites du travail et perspectives futures

https://www.projet-pfc.net/

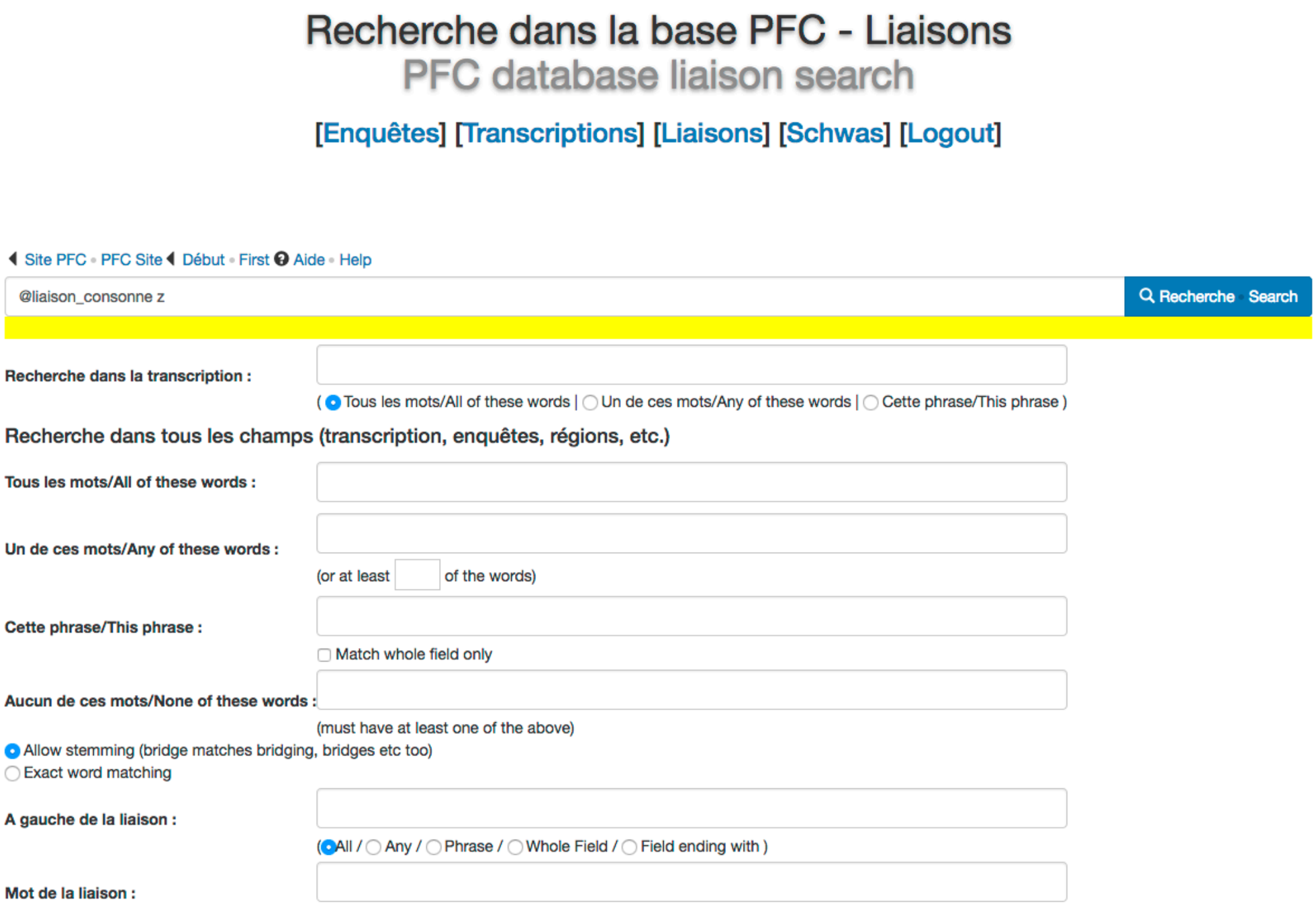
PFC (Phonologie du Français Contemporain) est un programme de recherche offrant une base de données de français oral contemporain dans l'espace francophone



La base de données
Durand, Laks & Lyche (2002)



L'interface


Recherche dans la base PFC - Liaisons
PFC database liaison search
[Enquêtes] [Transcriptions] [Liaisons] [Schwas] [Logout]
Site PFC • PFC Site Début • First Aide • Help
@liaison_consonne z
Recherche Search
Recherche dans la transcription :
(Tous les mots/All of these words | Un de ces mots/Any of these words | Cette phrase/This phrase)
Recherche dans tous les champs (transcription, enquêtes, régions, etc.)
Tous les mots/All of these words :
Un de ces mots/Any of these words :
(or at least of the words)
Cette phrase/This phrase :
Match whole field only
Aucun de ces mots/None of these words :
(must have at least one of the above)
Allow stemming (bridge matches bridging, bridges etc too)
Exact word matching
A gauche de la liaison :
(All / Any / Phrase / Whole Field / Field ending with)
Mot de la liaison :

3.1 Resultats de la recherche: @consonne z @ Dijon @ Sexe M

VILLE

ID LOCUTEUR

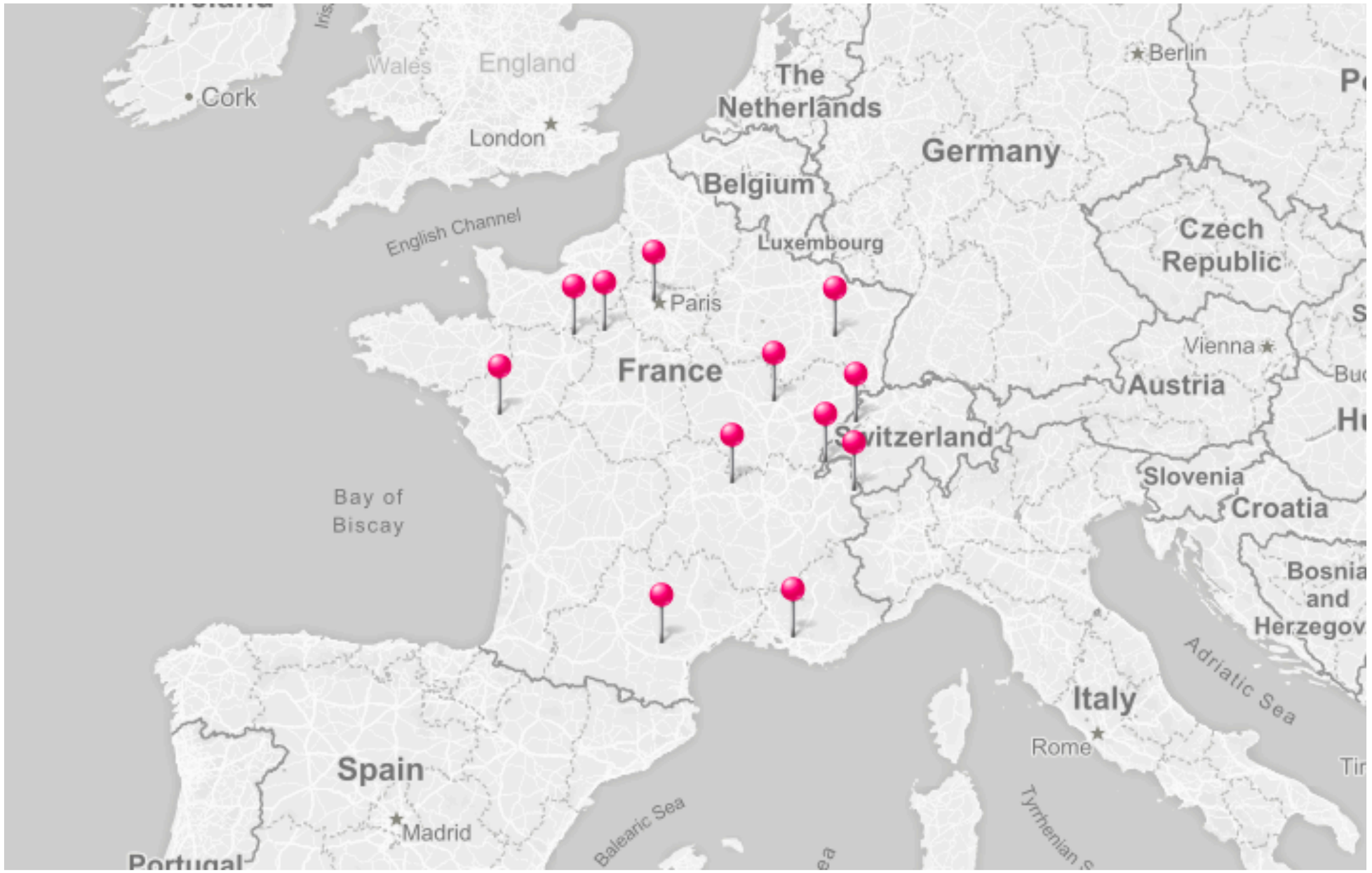
LIAISON RÉALISÉE

La recherche pour *@liaison_consonne z @loc_sexe m @enquete \"dijon\"* a trouvé 85 réponses

Exporter toutes les réponses au format CSV

Info	Locuteur	Enquête	Contexte Gauche	Liaison	Écoute	Contexte droit	L1	L2	L3	L4	L5
4035	21abl1	Dijon	quatrième aux jeux	jeux [z] olympiques		olympiques de Berlin en 1936 et plus récemment son usine de pâtes20 italiennes	1	z			
4042	21abl1	Dijon	Le maire de Beaulieu - Marc Blanc - est11t en revanche très	très [z] inquiet		inquiet	1	z			
4043	21abl1	Dijon	La cote du Premier Ministre ne cesse de baisser depuis les	les [z] élections		élections	1	z			
4046	21abl1	Dijon	La côte escarpée du Mont Saint-Pierre qui mène au village connaît des barrages chaque fois que les	les [z] opposants		opposants de tous les bords manifestent leur colère	1	z			
4053	21abl1	Dijon	ous	Nous [z] avons		avons le soutien du village entier	1	z			
4054	21abl1	Dijon	De plus quelques	quelques [z] articles		articles parus dans La Dépêche du Centre L'Express uest France et Le ouvel bservateur	1	z			
4055	21abl1	Dijon	indiqueraient que des	des [z] activistes		activistes des communes voisines préparent20 une journée chaude au Premier Ministre	1	z			
9900	21abl1	Dijon	BL: Ben c'était euh c'était hard quoi En plus le système quand tu tu prends10 une nounou tu as des	des [z] aides		aides mais qui tombent10 au bout du de trois mois	1	z			
9904	21abl1	Dijon	BL: les	les [z] aides		aides10 au logement euh le RMI on touche à peu près cinq mille et quelques francs cinq mille trois cent francs	1	z			
9911	21abl1	Dijon	BL: Bon après c'est vrai que on a quand c'était vraiment très galère on j'ai pu on j'ai pu avoir des	des [z] aides		aides quoi de ma mère quoi	1	z			

3.2 Points d'enquêtes (limités à la France) de PFC



3410	21aml1	Dijon	ML : Poser des plans tout le temps quoi tu vois une fois de temps	temps [z] en	en temps euh accord moi ça m'arrive aussi mais euh	1	z	
22100	44ajs2	Nantes	JS: Euh ça aggrave pas mais elle a pas récupéré mais par contre elle est	est [ ] en	en forme	0		

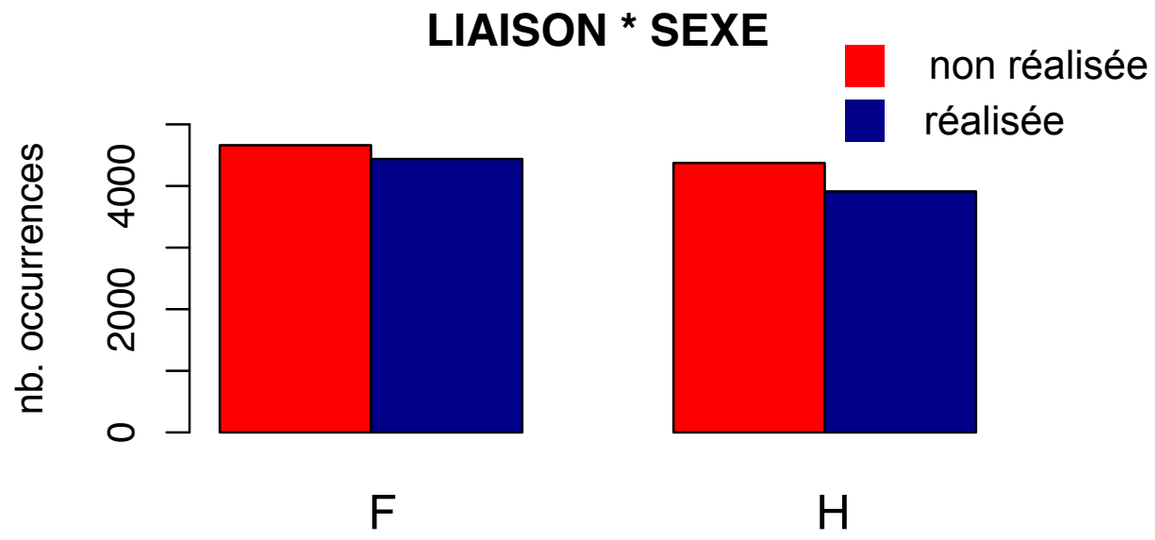
Détail dataset de PFC (après nettoyage semi-automatique)

LIAISON	LOCUTEUR	ÉDUCATION	SEXE	AGE	VILLE	MOT_GAUCHE	MOT_DROITE	POS_GAUCHE	POS_DROITE
OUI	21aml1	20	F	65	Dijon	temps	en	NOM	PRP
NON	44ajs2	14	F	59	Nantes	est	en	VER	PRP
…	…	…	…	…	…	…	…	…	…
oui = 8350 no = 9036 TOT = 17386	niveaux = 192	$\overline{\text{ÉDUCATION}}$ = 15,41	F = 9100 H = 8286	$\overline{\text{AGE}}$ = 48,11	niveaux = 20	niveaux = 487	niveaux = 487	niveaux = 11	niveaux = 7

- Objectif du travail
- La liaison - une petite présentation
- Un corpus phonologique : PFC
- Les outils statistiques (RL et GAM)
- Le model final et le résultats
- Limites du travail et perspectives futures

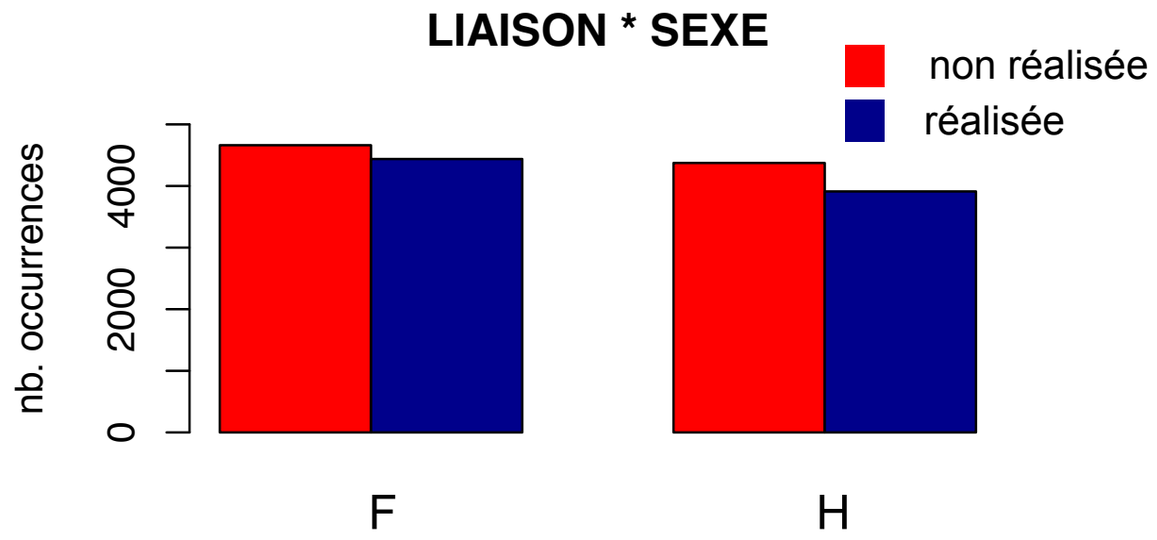
- Croiser deux variables catégorielles
- Tableaux de contingence LIAISON * SEXE

	F	H	Tot
Réalisée	4438	3912	8350
NON réalisée	4662	4374	9036
Tot	9100	8286	**17386**



- Croiser deux variables catégorielles
- Tableaux de contingence LIAISON * SEXE

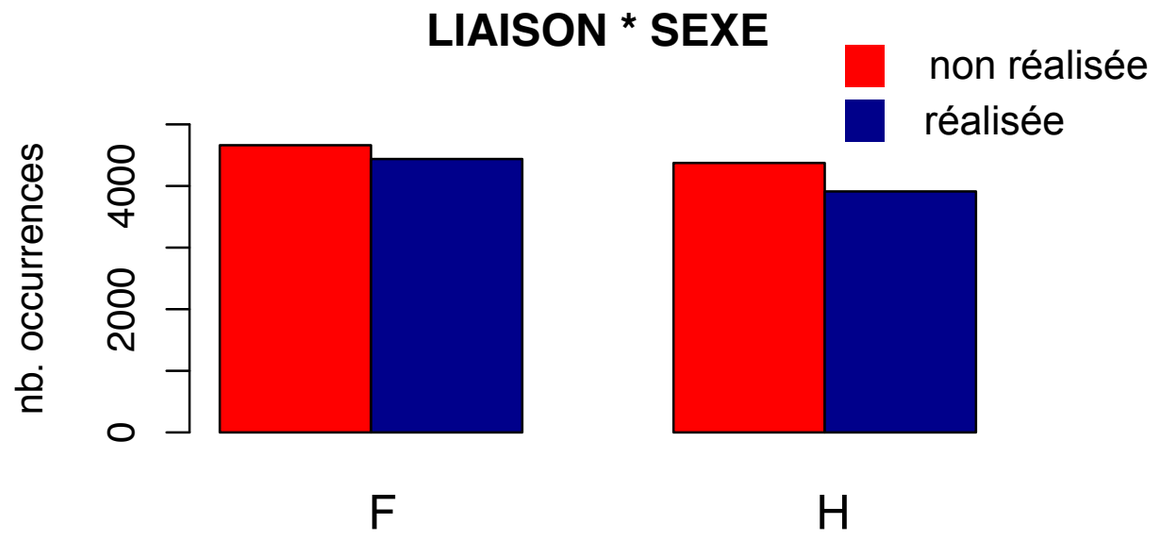
	F	H	Tot
Réalisée	4438	3912	8350
NON réalisée	4662	4374	9036
Tot	9100	8286	**17386**



Pour savoir si le sexe a une influence significative sur la réalisation/non-realisation de la liaison, nous pouvons effectuer un test d'indépendance du χ^2

- Croiser deux variables catégorielles
- Tableaux de contingence LIAISON * SEXE

	F	H	Tot
Réalisée	4438	3912	8350
NON réalisée	4662	4374	9036
Tot	9100	8286	**17386**

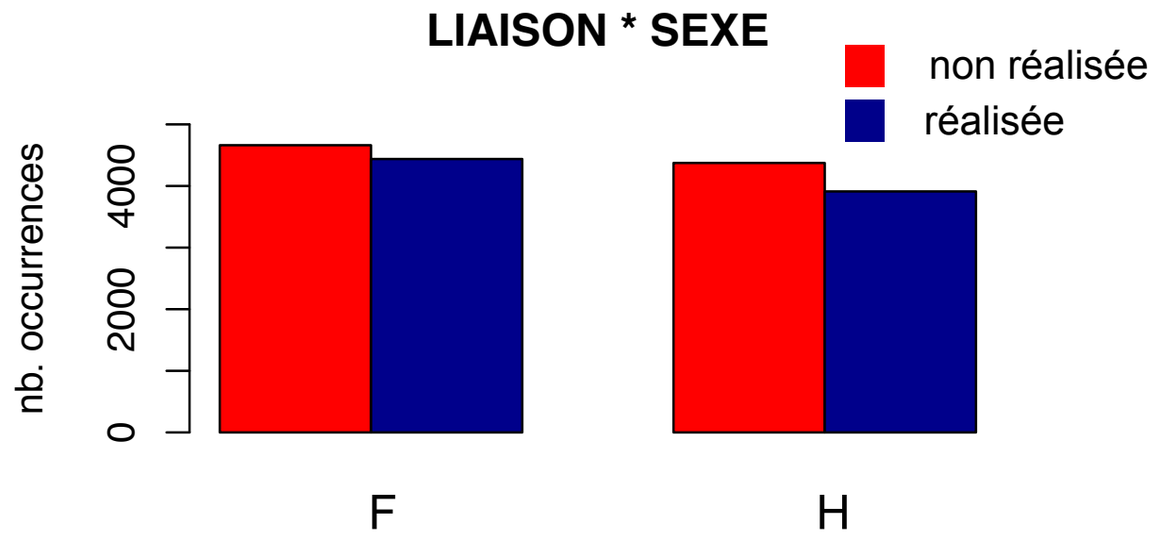


Pour savoir si le sexe a une influence significative sur la réalisation/non-realisation de la liaison, nous pouvons effectuer un test d'indépendance du χ^2

Le test est significatif : χ^2 (1) = 4.1503, *p-value* < 0.05*

- Croiser deux variables catégorielles
- Tableaux de contingence LIAISON * SEXE

	F	H	Tot
Réalisée	4438	3912	8350
NON réalisée	4662	4374	9036
Tot	9100	8286	**17386**



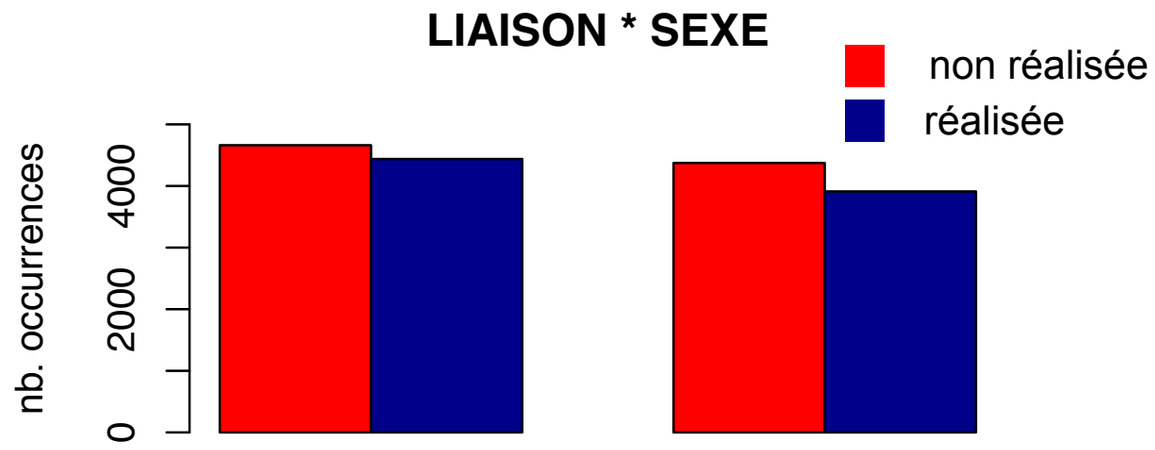
Pour savoir si le sexe a une influence significative sur la réalisation/non-realisation de la liaison, nous pouvons effectuer un test d'indépendance du χ^2

Le test est significatif : χ^2 (1) = 4.1503, *p-value* < 0.05*

Il existe une dependance significative entre le nombre de liaisons et le sexe

- Croiser deux variables catégorielles
- Tableaux de contingence LIAISON * SEXE

	F	H	Tot
Réalisée	4438	3912	8350
NON réalisée	4662	4374	9036
Tot	9100	8286	**17386**



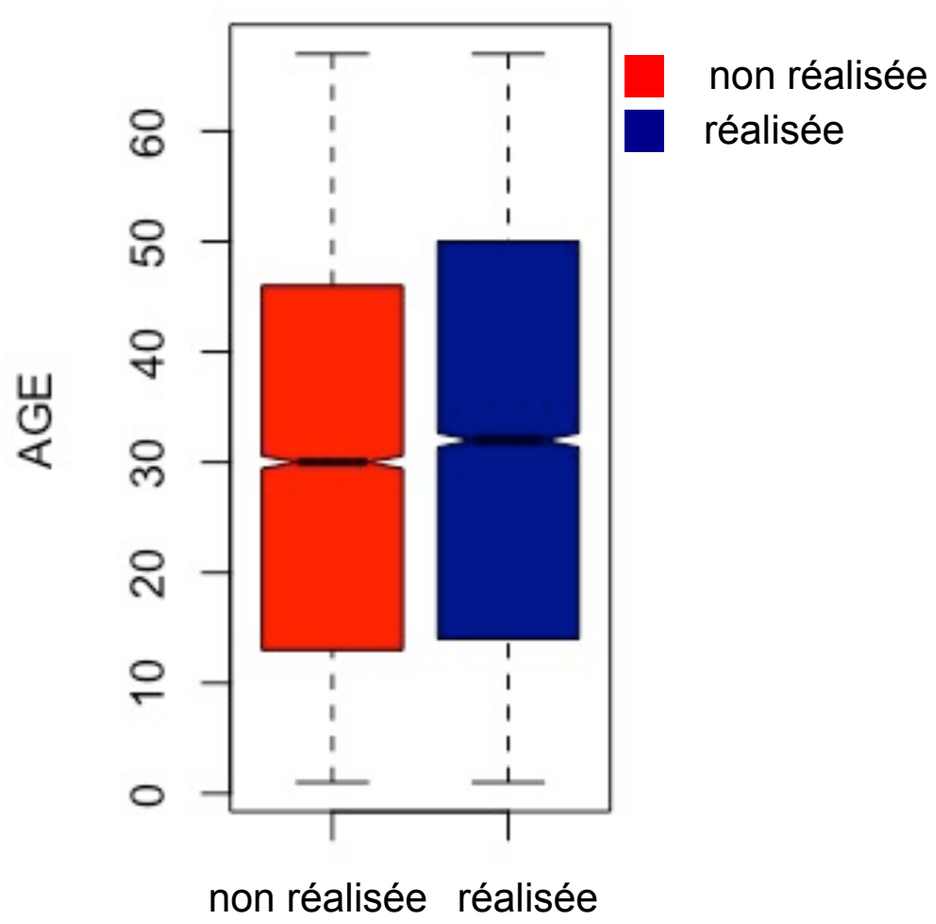
Pour savoir si le sexe a une influence significative sur la réalisation/non-realisation de la liaison, nous pouvons effectuer un test d'indépendance du χ^2

Le test est significatif : χ^2 (1) = 4.1503, *p-value* < 0.05*

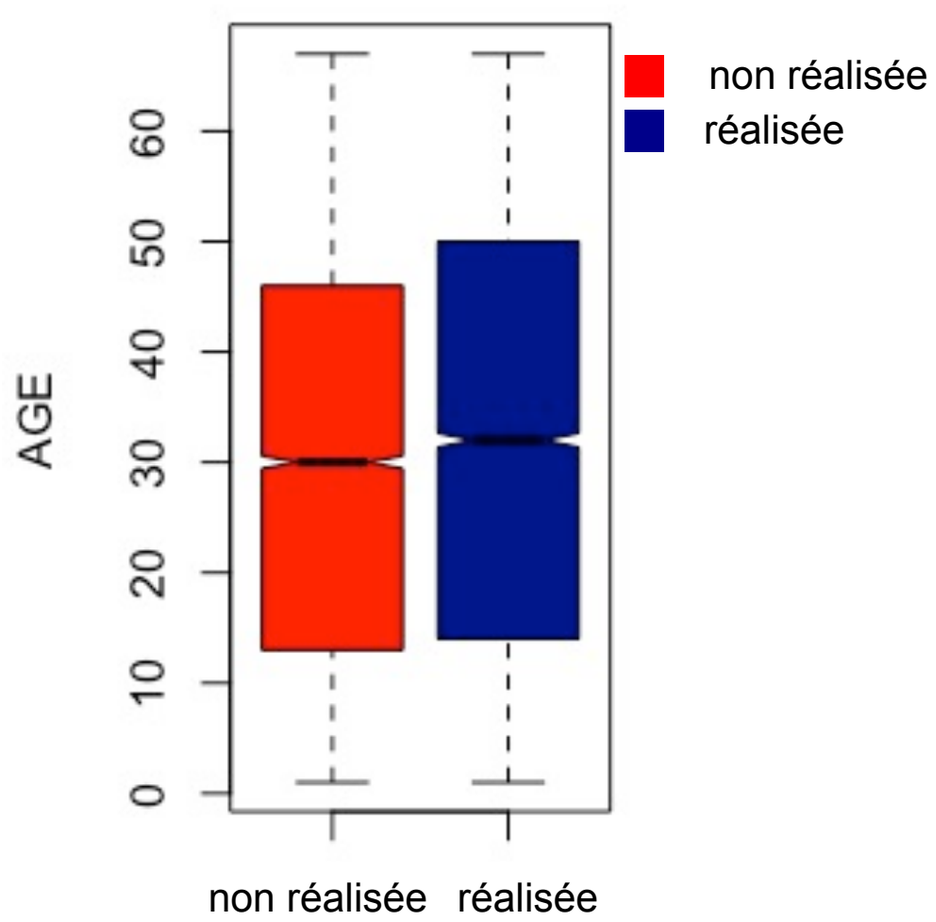
Il existe une dependance significative entre le nombre de liaisons et le sexe

C'est une test plutôt limitant et fortement lié aux données, qui ne permet pas une généralisation prédictive

- Croiser une variables catégorielle (LIAISON) avec une variable quantitative (AGE)

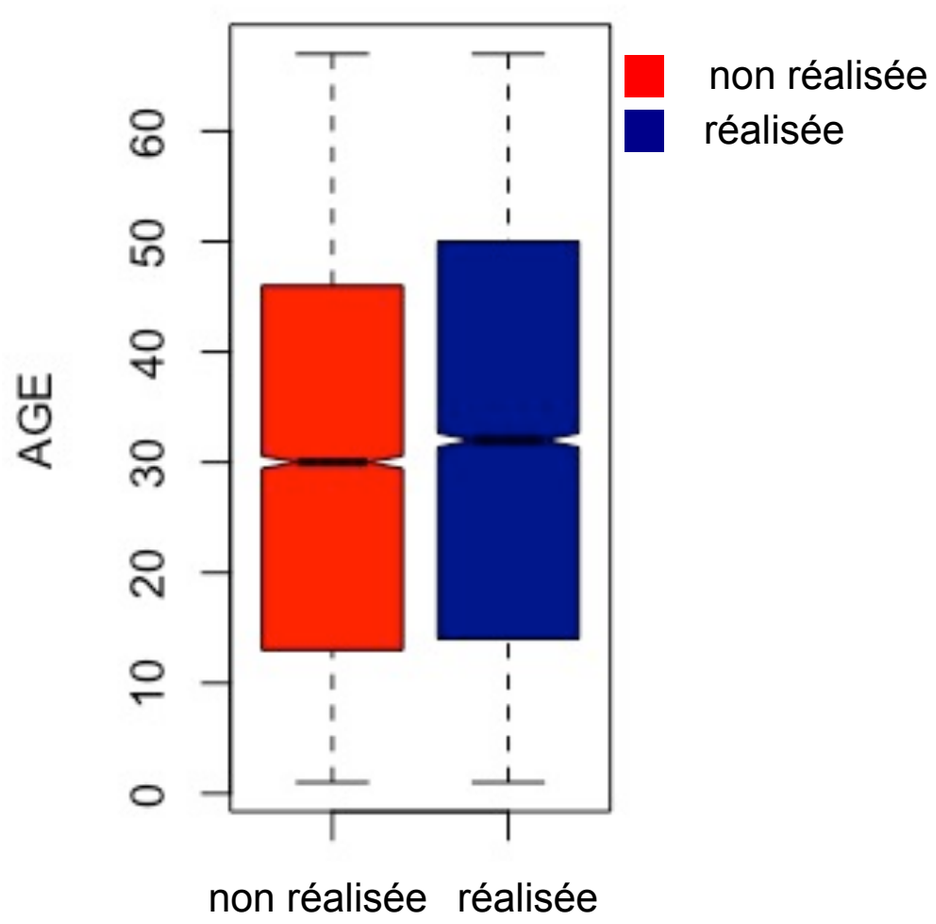


- Croiser une variables catégorielle (LIAISON) avec une variable quantitative (AGE)



Pour savoir si les deux groupes (réalisée vs. non réalisée) sont différents par rapport à l'âge des locuteurs, nous pouvons effectuer un test de variance (ANOVA)

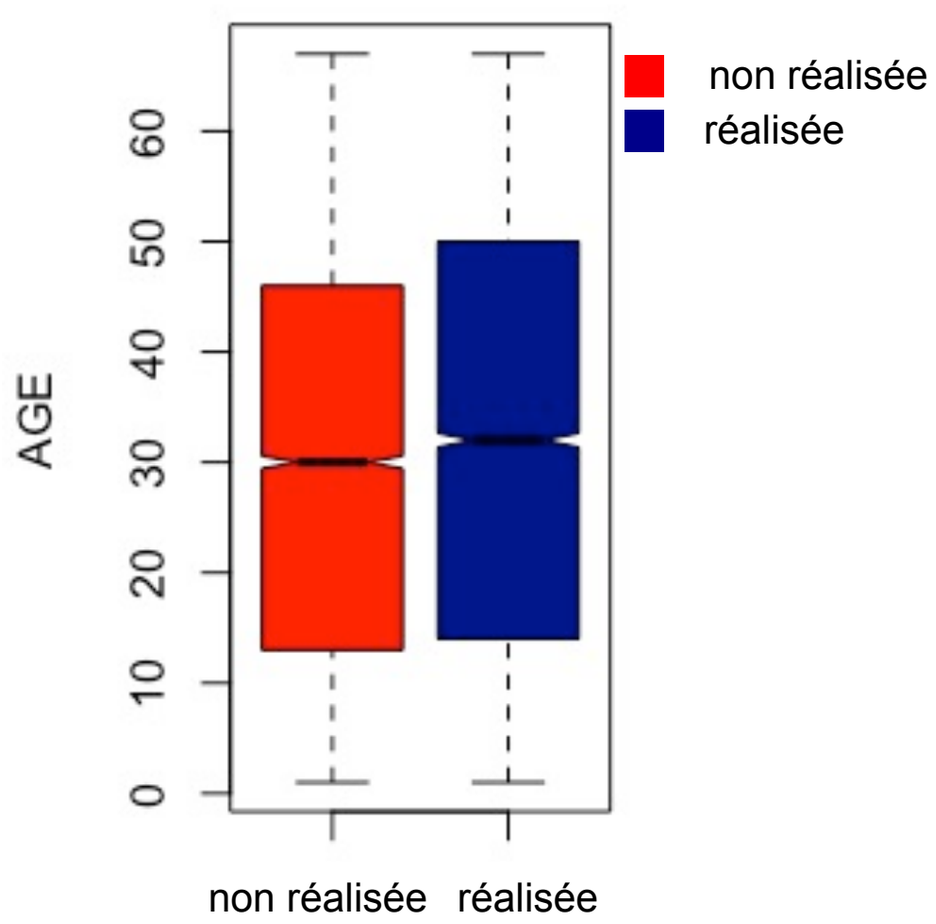
- Croiser une variables catégorielle (LIAISON) avec une variable quantitative (AGE)



Pour savoir si les deux groupes (réalisée vs. non réalisée) sont différents par rapport à l'âge des locuteurs, nous pouvons effectuer un test de variance (ANOVA)

Le test est significatif : ANOVA (1,15863) = 37.084, *p-value* < 0.001***

- Croiser une variables catégorielle (LIAISON) avec une variable quantitative (AGE)



Pour savoir si les deux groupes (réalisée vs. non réalisée) sont différents par rapport à l'âge des locuteurs, nous pouvons effectuer un test de variance (ANOVA)

Le test est significatif : ANOVA (1,15863) = 37.084, *p-value* < 0.001***

Les deux groupes sont statistiquement différents mais le test est limité à une statistique descriptive

Il nous faut une approche prédictive globale, capable de synthétiser dans un modèle unique toutes les variables à tester (qualitatives et quantitatives) et fournir des valeurs de significativité

Il nous faut une approche prédictive globale, capable de synthétiser dans un modèle unique toutes les variables à tester (qualitatives et quantitatives) et fournir des valeurs de significativité

Nécessité d'expliciter la nature de la relation entre variables (et pas seulement tester si deux variables sont différentes ou pas)

Il nous faut une approche prédictive globale, capable de synthétiser dans un modèle unique toutes les variables à tester (qualitatives et quantitatives) et fournir des valeurs de significativité

Nécessité d'expliciter la nature de la relation entre variables (et pas seulement tester si deux variables sont différentes ou pas)

Possibilité de faire de prédiction à partir des données analysées

4.1 La régression logistique (LR)

- La régression logistique (RL) permet de mesurer l'association entre la survenue (probabilité) d'un évènement (à expliquer) et les facteurs susceptibles de l'influencer (variables explicatives)

- La régression logistique (RL) permet de mesurer l'association entre la survenue (probabilité) d'un évènement (à expliquer) et les facteurs susceptibles de l'influencer (variables explicatives)

- Le but de la RL est de caractériser les relations entre la variable dépendante **Y** (ou variable à expliquer) et plusieurs variables $\mathbf{X_n}$ prises en compte simultanément (variables explicatives)

- La régression logistique (RL) permet de mesurer l'association entre la survenue (probabilité) d'un évènement (à expliquer) et les facteurs susceptibles de l'influencer (variables explicatives)

- Le but de la RL est de caractériser les relations entre la variable dépendante **Y** (ou variable à expliquer) et plusieurs variables $\mathbf{X_n}$ prises en compte simultanément (variables explicatives)

- La RL s'applique lorsque la variable à expliquer (Y) est qualitative (variable catégorielle avec nb. factors ≥ 2)

- La régression logistique (RL) permet de mesurer l'association entre la survenue (probabilité) d'un évènement (à expliquer) et les facteurs susceptibles de l'influencer (variables explicatives)

- Le but de la RL est de caractériser les relations entre la variable dépendante **Y** (ou variable à expliquer) et plusieurs variables $\mathbf{X_n}$ prises en compte simultanément (variables explicatives)

- La RL s'applique lorsque la variable à expliquer (Y) est qualitative (variable catégorielle avec nb. factors ≥ 2)

$$PLUIE_{OUI/NON} = NEBULOSITE$$

- La régression logistique (RL) permet de mesurer l'association entre la survenue (probabilité) d'un évènement (à expliquer) et les facteurs susceptibles de l'influencer (variables explicatives)

- Le but de la RL est de caractériser les relations entre la variable dépendante **Y** (ou variable à expliquer) et plusieurs variables $\mathbf{X_n}$ prises en compte simultanément (variables explicatives)

- La RL s'applique lorsque la variable à expliquer (Y) est qualitative (variable catégorielle avec nb. factors ≥ 2)

$$PLUIE_{OUI/NON} = NEBULOSITE$$

$$REUSSIR_{OUI/NON} = TEMPS - TRAVAIL$$

- La régression logistique (RL) permet de mesurer l'association entre la survenue (probabilité) d'un évènement (à expliquer) et les facteurs susceptibles de l'influencer (variables explicatives)

- Le but de la RL est de caractériser les relations entre la variable dépendante **Y** (ou variable à expliquer) et plusieurs variables $\mathbf{X_n}$ prises en compte simultanément (variables explicatives)

- La RL s'applique lorsque la variable à expliquer (Y) est qualitative (variable catégorielle avec nb. factors ≥ 2)

$$PLUIE_{OUI/NON} = NEBULOSITE$$

$$REUSSIR_{OUI/NON} = TEMPS - TRAVAIL$$

$$LIAISON_{OUI/NON} = variables X_n$$

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

La variable à expliquer (Y) est exprimée en fonction d'un intercept (ou ordonnée à l'origine) β_0 , d'une variable explicative (α_1) rattachée à son coefficient β_1 et d'un terme de bruit $\boldsymbol{\epsilon}$

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

La variable à expliquer (Y) est exprimée en fonction d'un intercept (ou ordonnée à l'origine) β_0 , d'une variable explicative (α_1) rattachée à son coefficient β_1 et d'un terme de bruit $\boldsymbol{\epsilon}$

Y est la variable catégorielle à expliquer : $\text{PLUIE}_{\text{OUI/NON}}$, $\text{REUSSIR}_{\text{OUI/NON}}$

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

La variable à expliquer (Y) est exprimée en fonction d'un intercept (ou ordonnée à l'origine) β_0 , d'une variable explicative (α_1) rattachée à son coefficient β_1 et d'un terme de bruit $\boldsymbol{\epsilon}$

Y est la variable catégorielle à expliquer : $\text{PLUIE}_{\text{OUI/NON}}$, $\text{REUSSIR}_{\text{OUI/NON}}$

Le paramètre $\boldsymbol{\alpha_1}$ représente la variable explicative du modèle: NÉBULOSITÉ, TEMPS_DE_TRAVAIL

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

La variable à expliquer (Y) est exprimée en fonction d'un intercept (ou ordonnée à l'origine) β_0 , d'une variable explicative (α_1) rattachée à son coefficient β_1 et d'un terme de bruit $\boldsymbol{\epsilon}$

Y est la variable catégorielle à expliquer : $PLUIE_{OUI/NON}$, $REUSSIR_{OUI/NON}$

Le paramètre $\boldsymbol{\alpha_1}$ représente la variable explicative du modèle: NÉBULOSITÉ, TEMPS_DE_TRAVAIL

Le coefficient $\boldsymbol{\beta_1}$ mesure l'association entre la variable explicative (α_1) et la variable expliquée (Y). Il correspond au logarithme de l'odds ratio β=log (OR)

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

La variable à expliquer (Y) est exprimée en fonction d'un intercept (ou ordonnée à l'origine) β_0 , d'une variable explicative (α_1) rattachée à son coefficient β_1 et d'un terme de bruit $\boldsymbol{\epsilon}$

Y est la variable catégorielle à expliquer : $PLUIE_{OUI/NON}$, $REUSSIR_{OUI/NON}$

Le paramètre $\boldsymbol{\alpha_1}$ représente la variable explicative du modèle: NÉBULOSITÉ, TEMPS_DE_TRAVAIL

Le coefficient $\boldsymbol{\beta_1}$ mesure l'association entre la variable explicative (α_1) et la variable expliquée (Y). Il correspond au logarithme de l'odds ratio β=log (OR)

L'intercept $\boldsymbol{\beta_0}$ correspond aux valeurs baseline du modèle lorsque les variables explicatives sont égales à 0 (β_n=0).

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

La variable à expliquer (Y) est exprimée en fonction d'un intercept (ou ordonnée à l'origine) β_0 , d'une variable explicative (α_1) rattachée à son coefficient β_1 et d'un terme de bruit $\boldsymbol{\epsilon}$

Y est la variable catégorielle à expliquer : $PLUIE_{OUI/NON}$, $REUSSIR_{OUI/NON}$

Le paramètre $\boldsymbol{\alpha_1}$ représente la variable explicative du modèle: NÉBULOSITÉ, TEMPS_DE_TRAVAIL

Le coefficient $\boldsymbol{\beta_1}$ mesure l'association entre la variable explicative (α_1) et la variable expliquée (Y). Il correspond au logarithme de l'odds ratio β=log (OR)

L'intercept $\boldsymbol{\beta_0}$ correspond aux valeurs baseline du modèle lorsque les variables explicatives sont égales à 0 (β_n=0).

$\boldsymbol{\epsilon}$ est un terme d'erreur non contrôlé qui doit impérativement suivre une distribution normale et de même variance. C'est un paramètre négligeable

4.2 La régression logistique - définition mathématique

Dans le cas de nos données sur la liaison, on essaie de trouver un relation entre l'AGE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

Dans le cas de nos données sur la liaison, on essaie de trouver un relation entre l'AGE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

$$LOGIT[P(Liaison = oui) \mid \text{AGE})] = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

Dans le cas de nos données sur la liaison, on essaie de trouver un relation entre l'AGE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

$$LOGIT[P(Liaison = oui) \mid \text{AGE})] = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

<u>Trouver les coefficients:</u>

β_0(Intercept)	-0.2713395
β_1	0.0043476

Dans le cas de nos données sur la liaison, on essaie de trouver un relation entre l'AGE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

$$LOGIT[P(Liaison = oui) \mid \text{AGE})] = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

<u>Trouver les coefficients:</u>

$\boldsymbol{\beta_0}$(Intercept)	-0.2713395
$\boldsymbol{\beta_1}$	0.0043476

β_1 = odds ratio de la probabilité d'avoir *Liaison=oui* quand AGE augmente d'une unité

Dans le cas de nos données sur la liaison, on essaie de trouver un relation entre l'AGE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

$$LOGIT[P(Liaison = oui) \mid \text{AGE})] = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

<u>Trouver les coefficients:</u>

β_0(Intercept)	-0.2713395
β_1	0.0043476

β_1 = odds ratio de la probabilité d'avoir *Liaison=oui* quand AGE augmente d'une unité

Regarder le signe : <u>POSITIF</u>

Dans le cas de nos données sur la liaison, on essaie de trouver un relation entre l'AGE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

$$LOGIT[P(Liaison = oui) \mid \text{AGE})] = \beta_0 + \beta_1 \text{AGE} + \epsilon$$

<u>Trouver les coefficients:</u>

β_0(Intercept)	-0.2713395
β_1	0.0043476

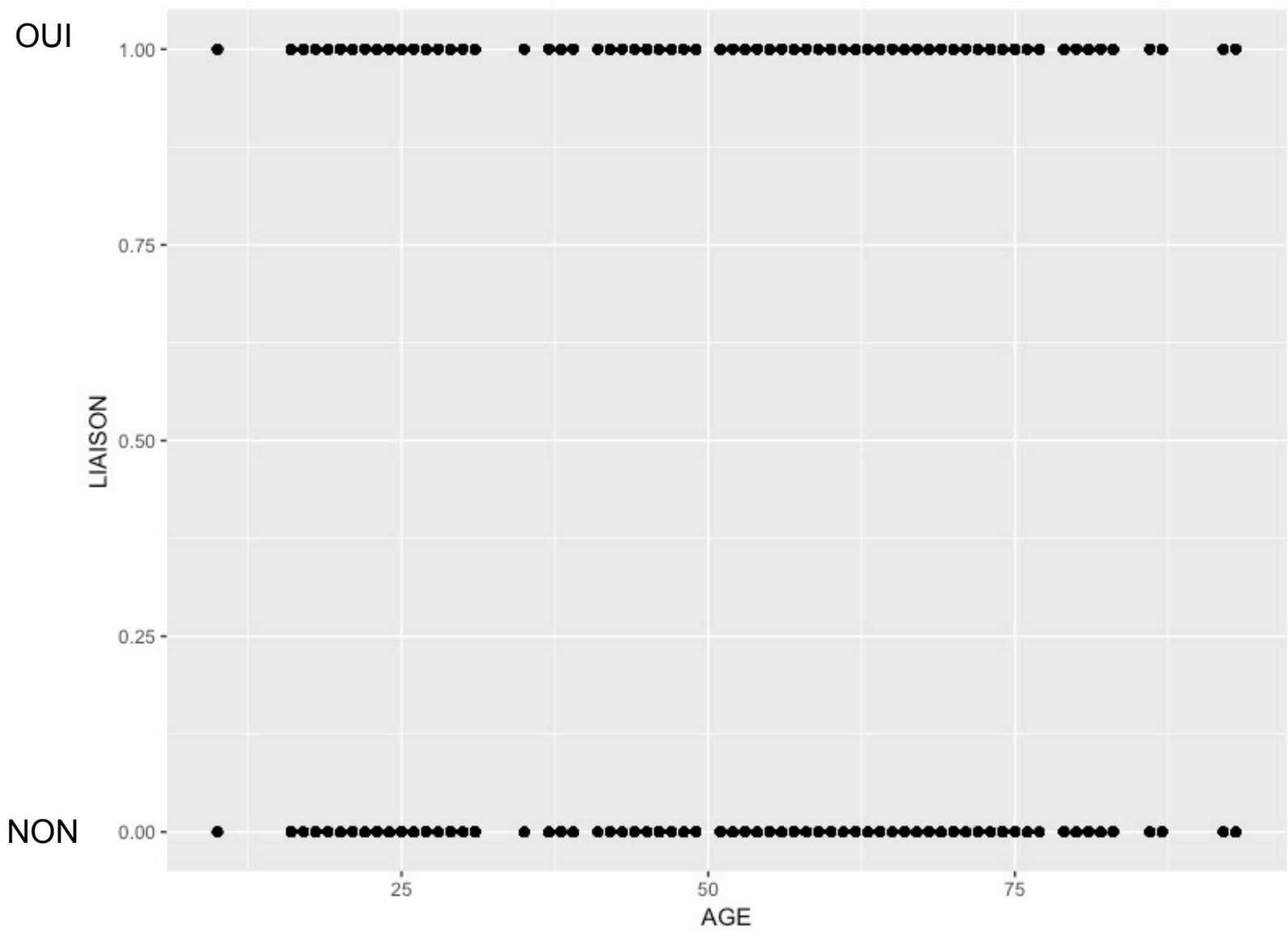
β_1 = odds ratio de la probabilité d'avoir *Liaison=oui* quand AGE augmente d'une unité

Regarder le signe : <u>POSITIF</u>

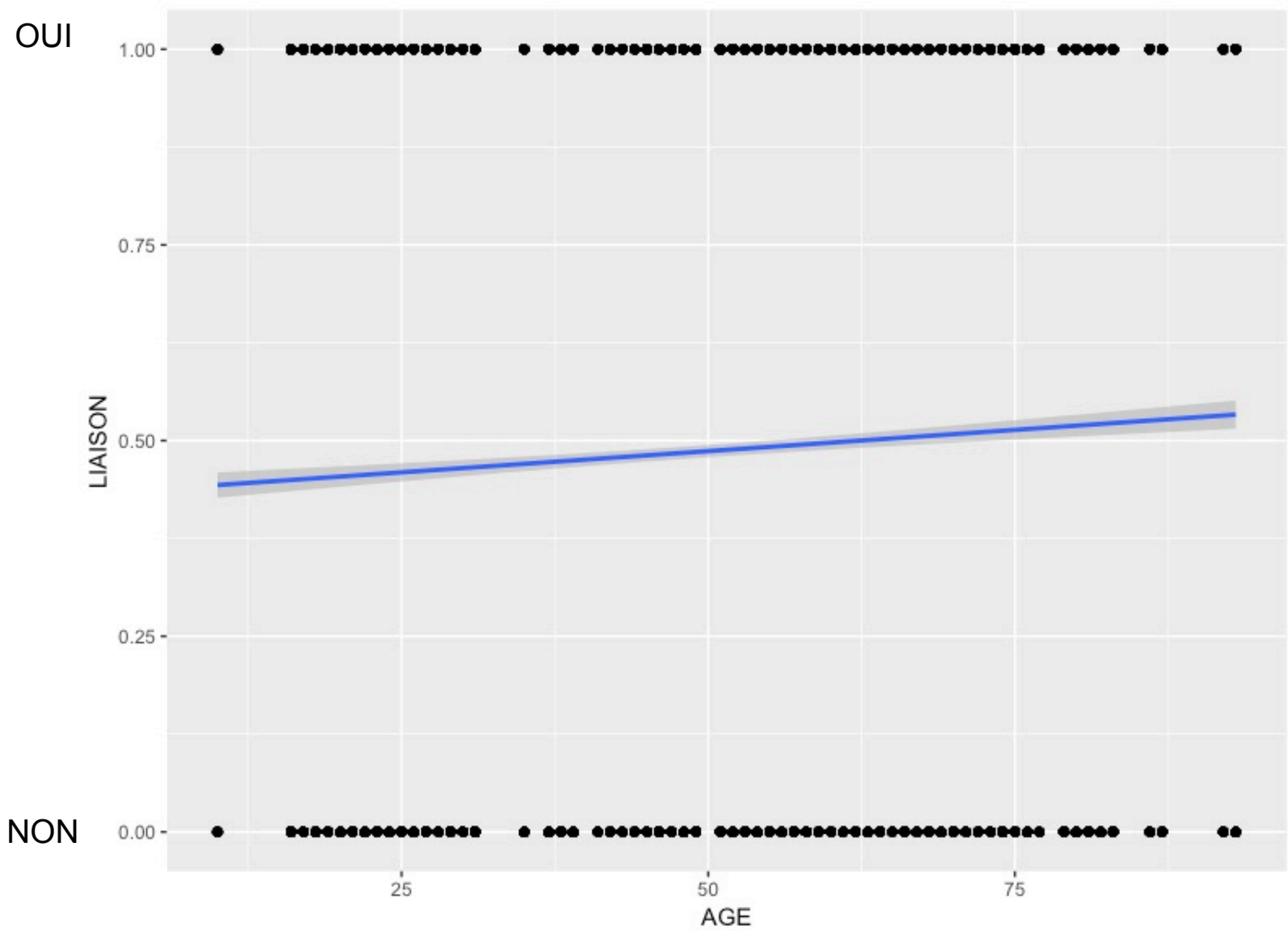
$$LOGIT[P(Liaison = oui \mid AGE] = -0.2713395 + 0.0043476 \text{ AGE} + \epsilon$$

A partir de cette équation on peut traces la fonction de régression *Liaison* ~ *AGE*

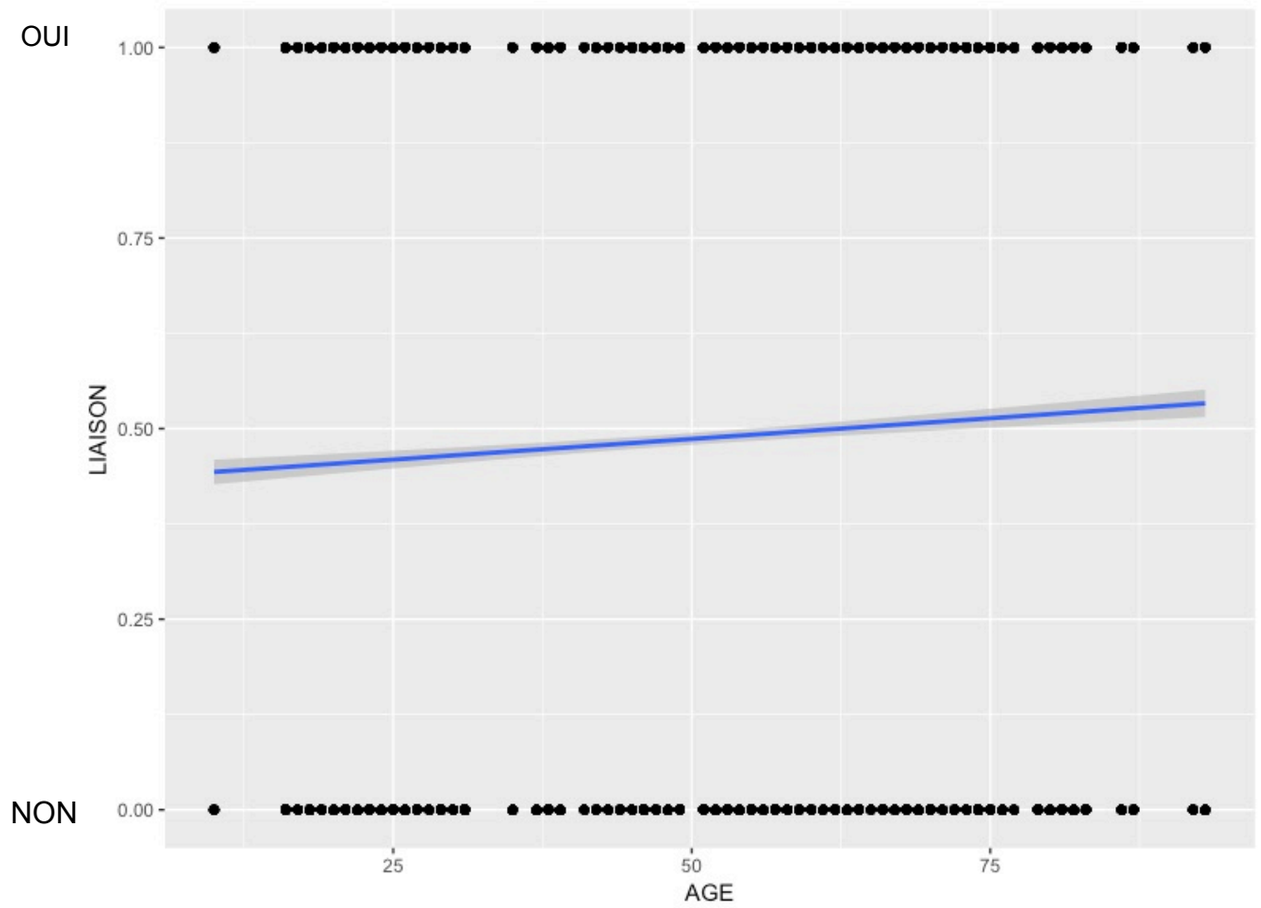
$$LOGIT[P(Liaison = oui \mid AGE] = -0.2713395 + 0.0043476 \text{ AGE} + \epsilon$$



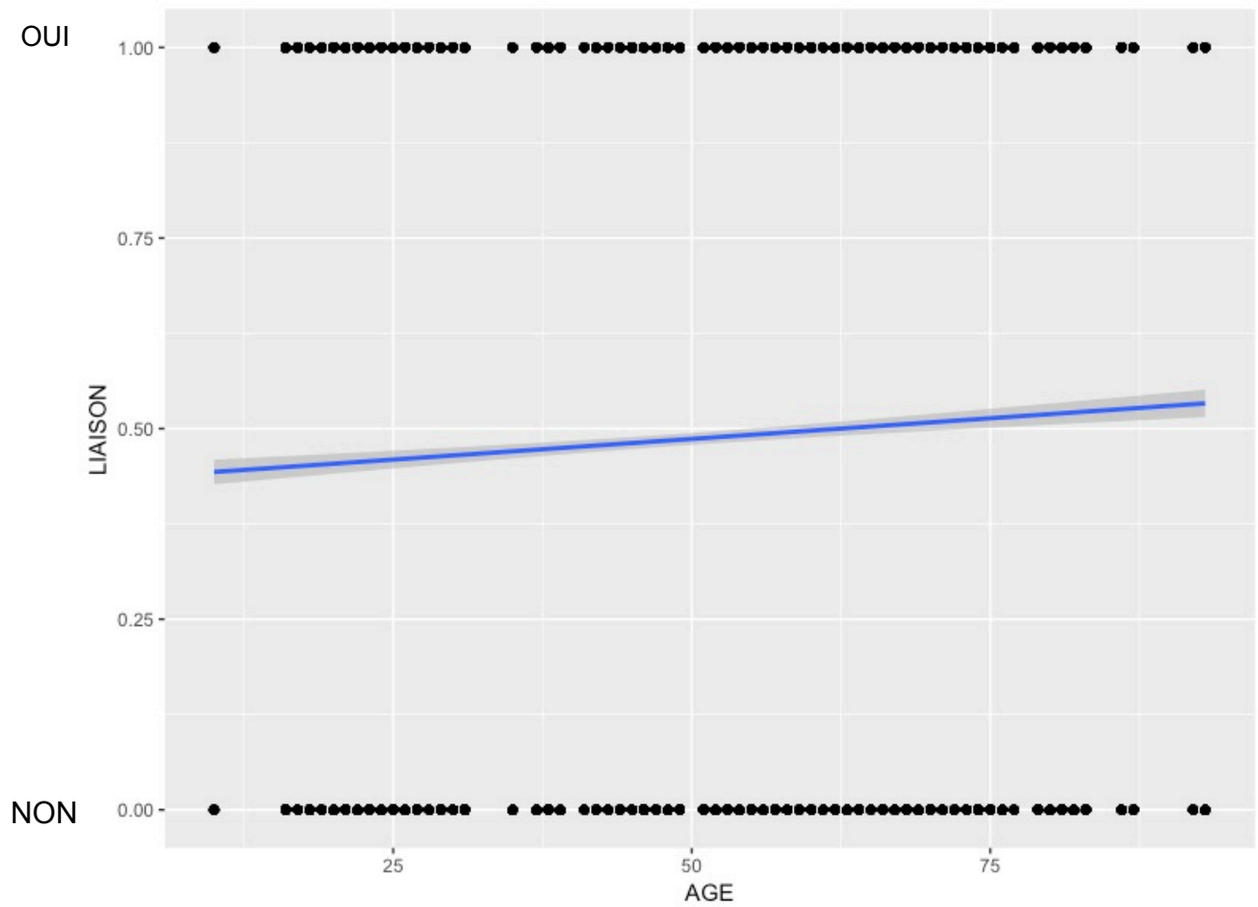
$$LOGIT[P(Liaison = oui \mid AGE] = -0.2713395 + 0.0043476 \text{ AGE} + \epsilon$$



L'augmentation de l'AGE détermine une plus forte probabilité d'avoir une liaison réalisée

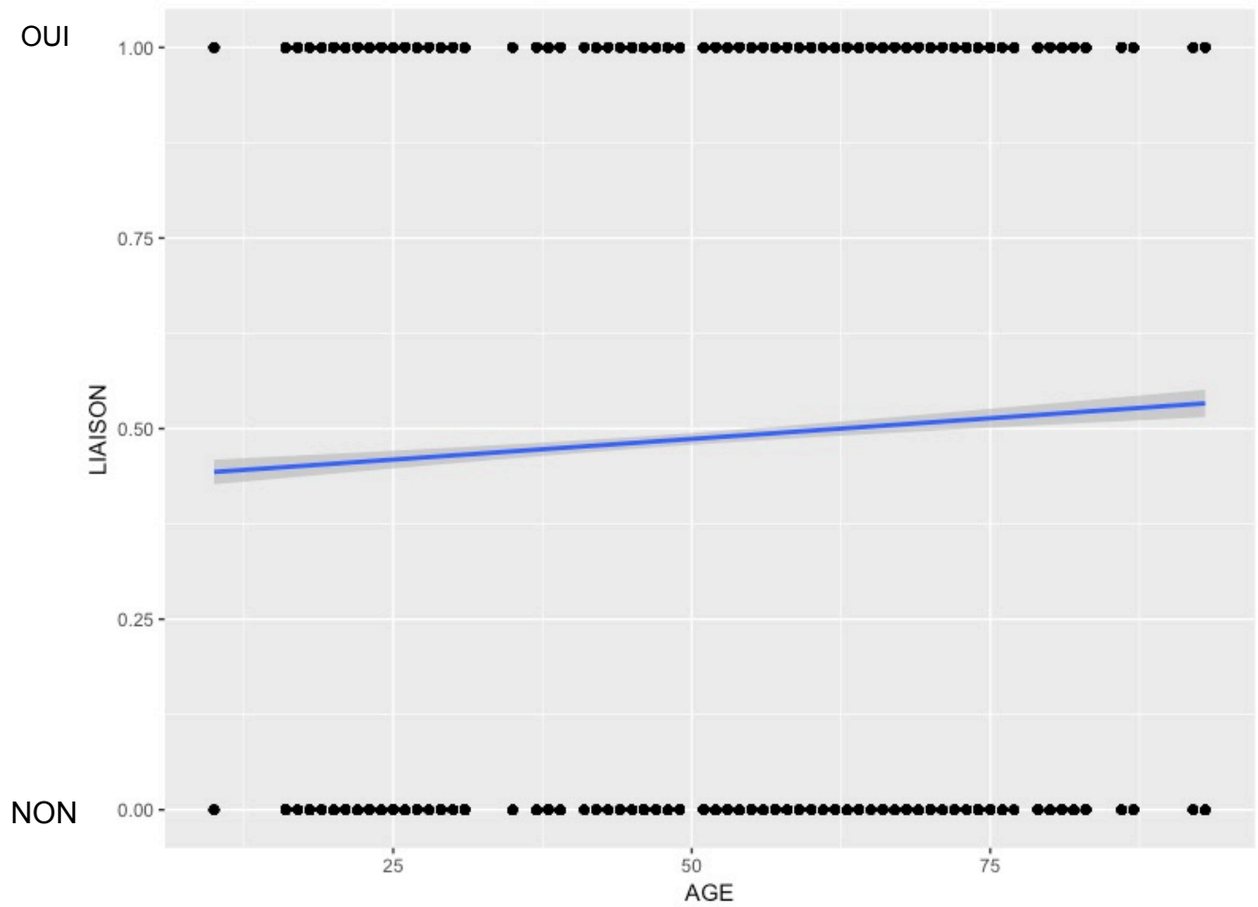


L'augmentation de l'AGE détermine une plus forte probabilité d'avoir une liaison réalisée



Deux mesures d'évaluation d'un model logistique:

L'augmentation de l'AGE détermine une plus forte probabilité d'avoir une liaison réalisée

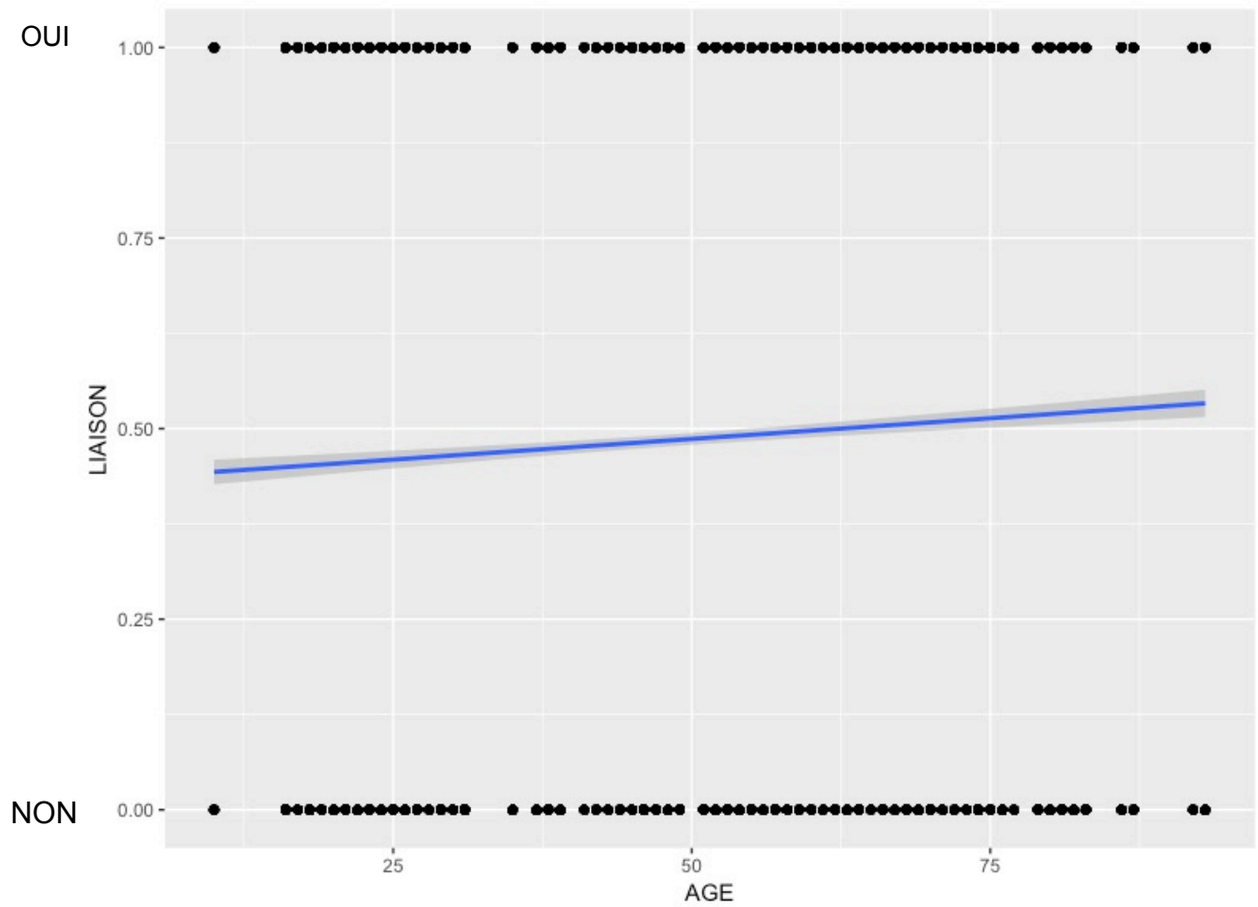


Deux mesures d'évaluation d'un model logistique:

1- **AIC**, critère d'information d'Akaike est une mesure de la qualité d'un modèle statistique

$AIC=2k-2\log(L)$ où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle. On choisit le modèle avec le AIC le plus faible

L'augmentation de l'AGE détermine une plus forte probabilité d'avoir une liaison réalisée



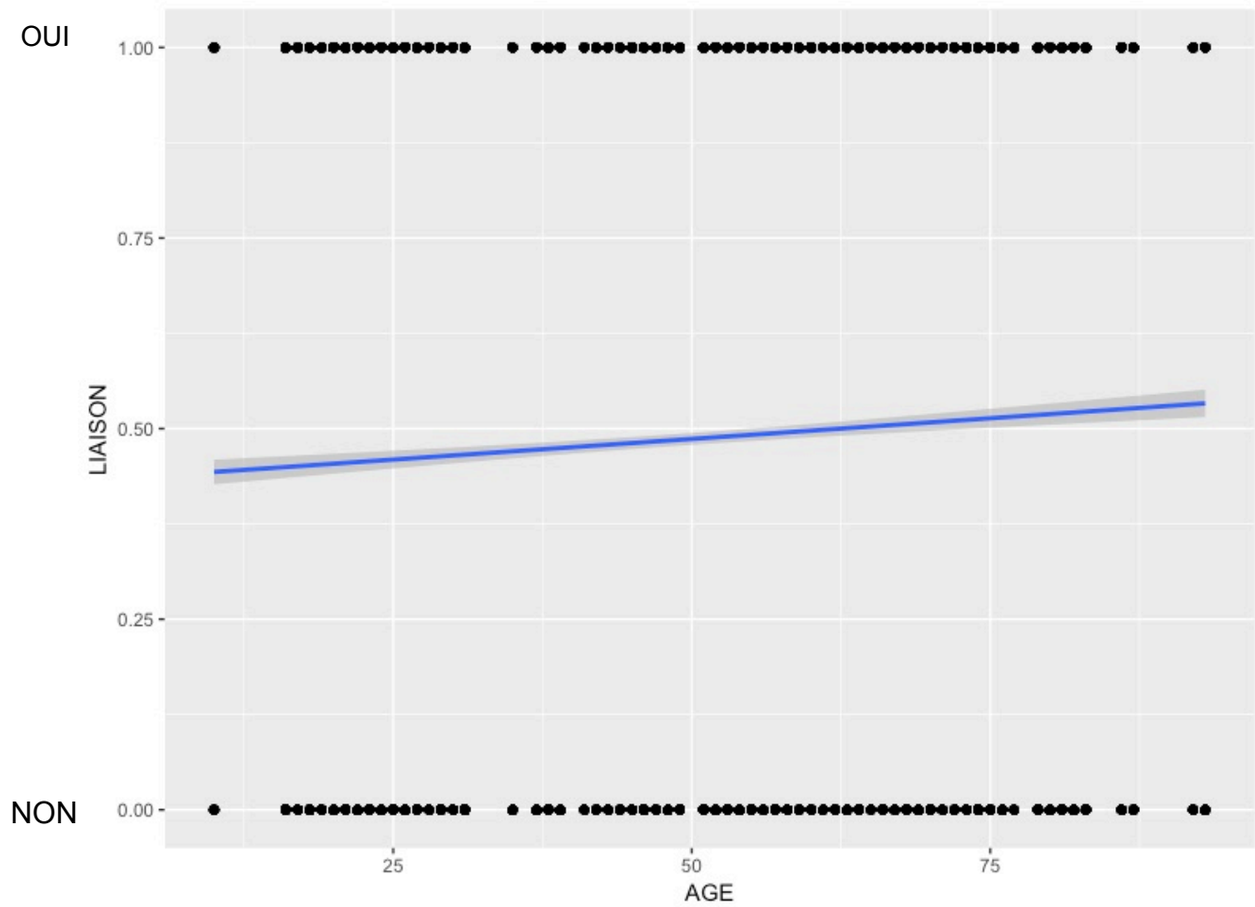
Deux mesures d'évaluation d'un model logistique:

1- **AIC**, critère d'information d'Akaike est une mesure de la qualité d'un modèle statistique

$AIC=2k-2\log(L)$ où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle. On choisit le modèle avec le AIC le plus faible

2- **La déviance expliquée du modèle (%)**, est une mesure de la qualité de la prédiction de la régression et représente la variabilité de la variable dépendante Y décrite par le modèle

L'augmentation de l'AGE détermine une plus forte probabilité d'avoir une liaison réalisée



AIC = 21947.7

Déviance expliquée = 0.15%

Deux mesures d'évaluation d'un model logistique:

1- **AIC**, critère d'information d'Akaike est une mesure de la qualité d'un modèle statistique

$AIC=2k-2\log(L)$ où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle. On choisit le modèle avec le AIC le plus faible

2- **La déviance expliquée du modèle (%)**, est une mesure de la qualité de la prédiction de la régression et représente la variabilité de la variable dépendante Y décrite par le modèle

4.3.1 Un premier modèle: Liaison ~ AGE- predictions

Probabilité de réaliser une liaison par une personne de 50 ans

$$\text{AGE} = 50$$

Probabilité de réaliser une liaison par une personne de 50 ans

$$\text{AGE} = 50$$

$$\text{P}(Liaison = oui \mid \text{AGE}) = \frac{\exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}{1 + \exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}$$

Probabilité de réaliser une liaison par une personne de 50 ans

$$\text{AGE} = 50$$

$$\text{P}(Liaison = oui \mid \text{AGE}) = \frac{\exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}{1 + \exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}$$

$$\text{P}(Liaison = oui \mid AGE = 50) = \frac{\exp(-0.2713395 + 0.0043476\ 50) + \epsilon}{1 + \exp(-0.2713395 + 0.0043476\ 50) + \epsilon}$$

Probabilité de réaliser une liaison par une personne de 50 ans

$$\text{AGE} = 50$$

$$\text{P}(Liaison = oui \mid \text{AGE}) = \frac{\exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}{1 + \exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}$$

$$\text{P}(Liaison = oui \mid AGE = 50) = \frac{\exp(-0.2713395 + 0.0043476\ 50) + \epsilon}{1 + \exp(-0.2713395 + 0.0043476\ 50) + \epsilon}$$

$$P(Liaison = oui \mid AGE = 50) = 0.491$$

Probabilité de réaliser une liaison par une personne de 50 ans

$$\text{AGE} = 50$$

$$\text{P}(Liaison = oui \mid \text{AGE}) = \frac{\exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}{1 + \exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}$$

$$\text{P}(Liaison = oui \mid AGE = 50) = \frac{\exp(-0.2713395 + 0.0043476\ 50) + \epsilon}{1 + \exp(-0.2713395 + 0.0043476\ 50) + \epsilon}$$

$$P(Liaison = oui \mid AGE = 50) = 0.491$$

$$P(Liaison = oui \mid AGE = 20) = 0.462$$

Probabilité de réaliser une liaison par une personne de 50 ans

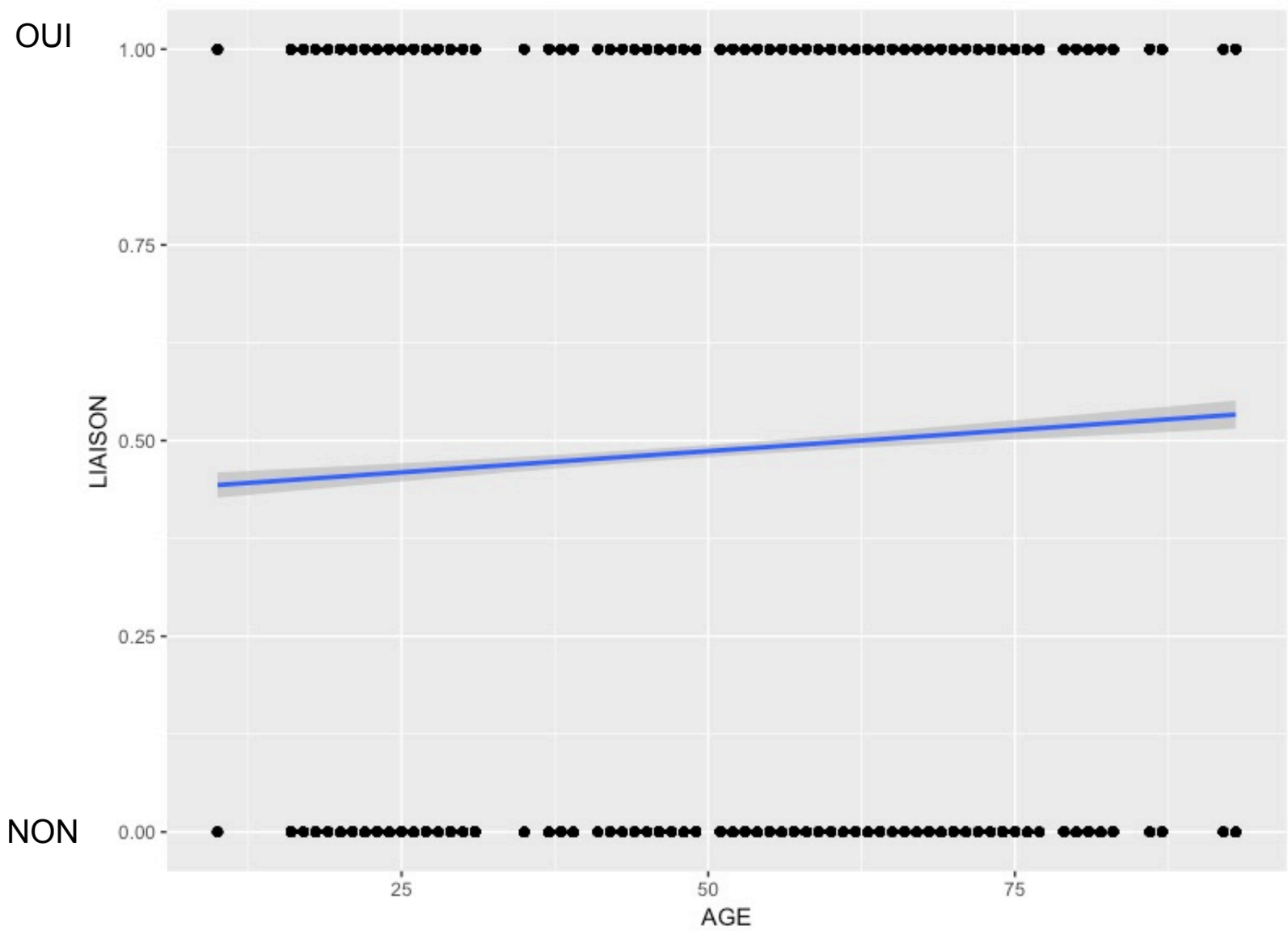
$$\text{AGE} = 50$$

$$\text{P}(Liaison = oui \mid \text{AGE}) = \frac{\exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}{1 + \exp(\beta_0 + \beta_1 \text{AGE} + \epsilon)}$$

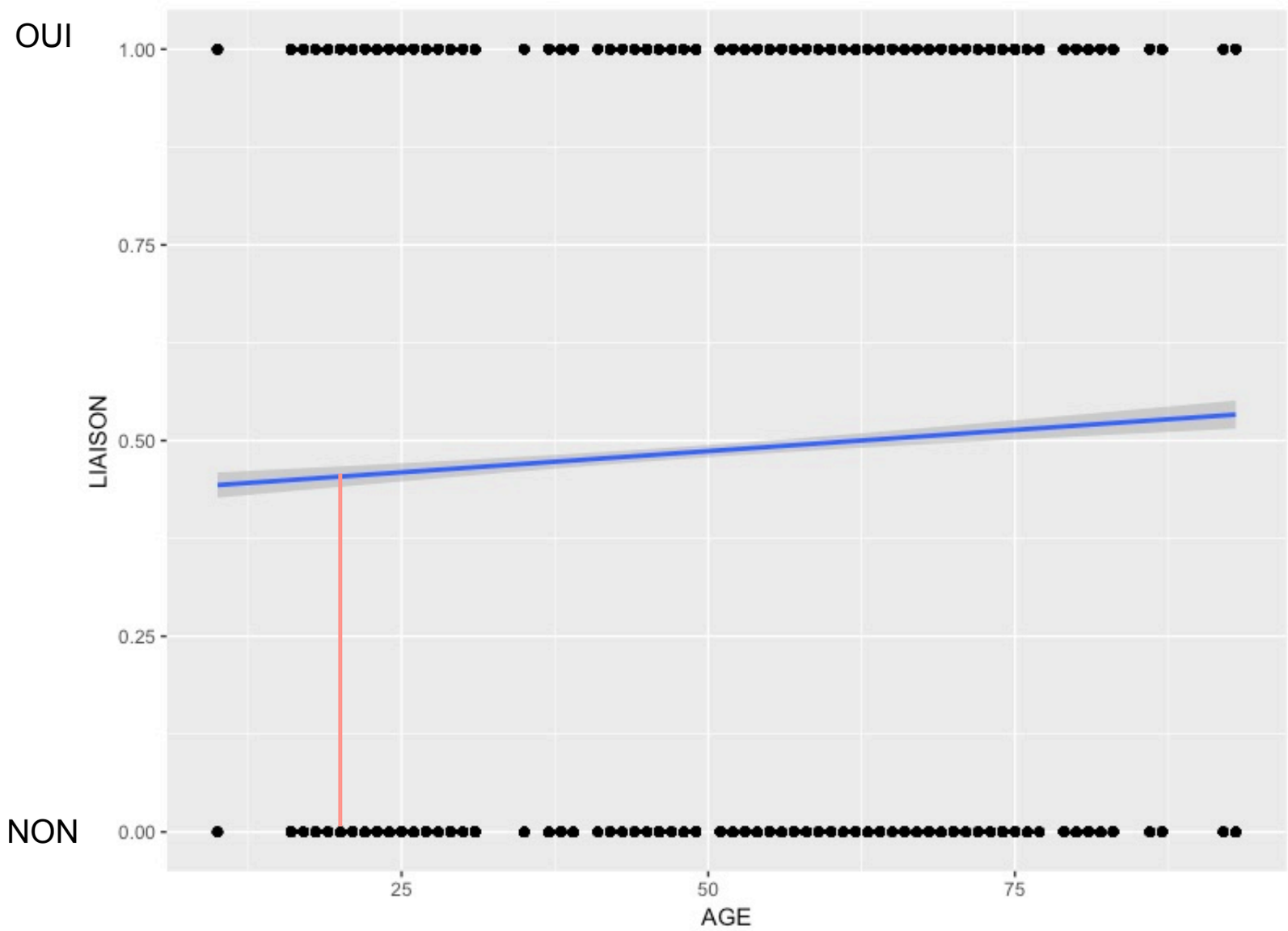
$$\text{P}(Liaison = oui \mid AGE = 50) = \frac{\exp(-0.2713395 + 0.0043476\ 50) + \epsilon}{1 + \exp(-0.2713395 + 0.0043476\ 50) + \epsilon}$$

$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$

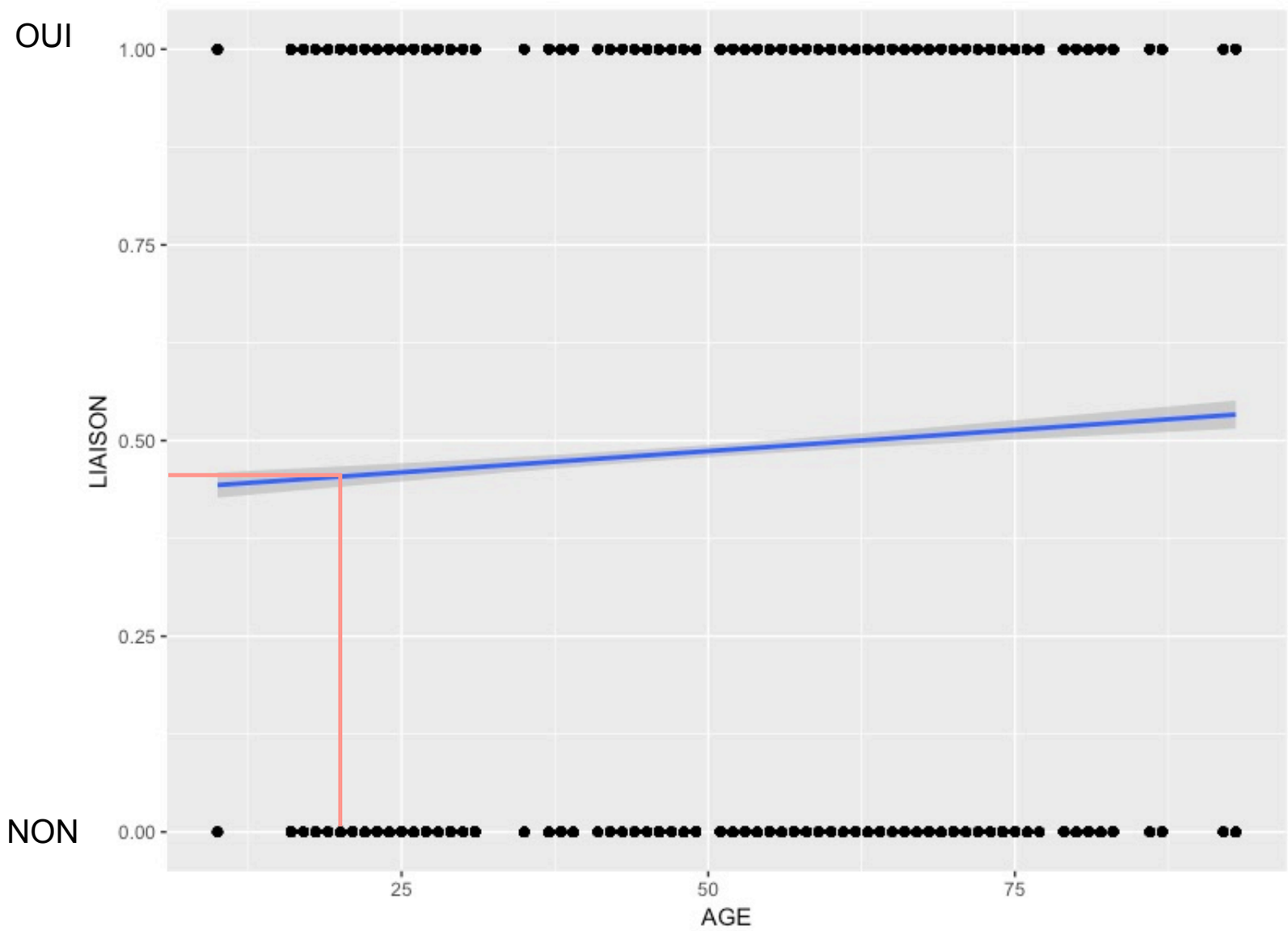
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



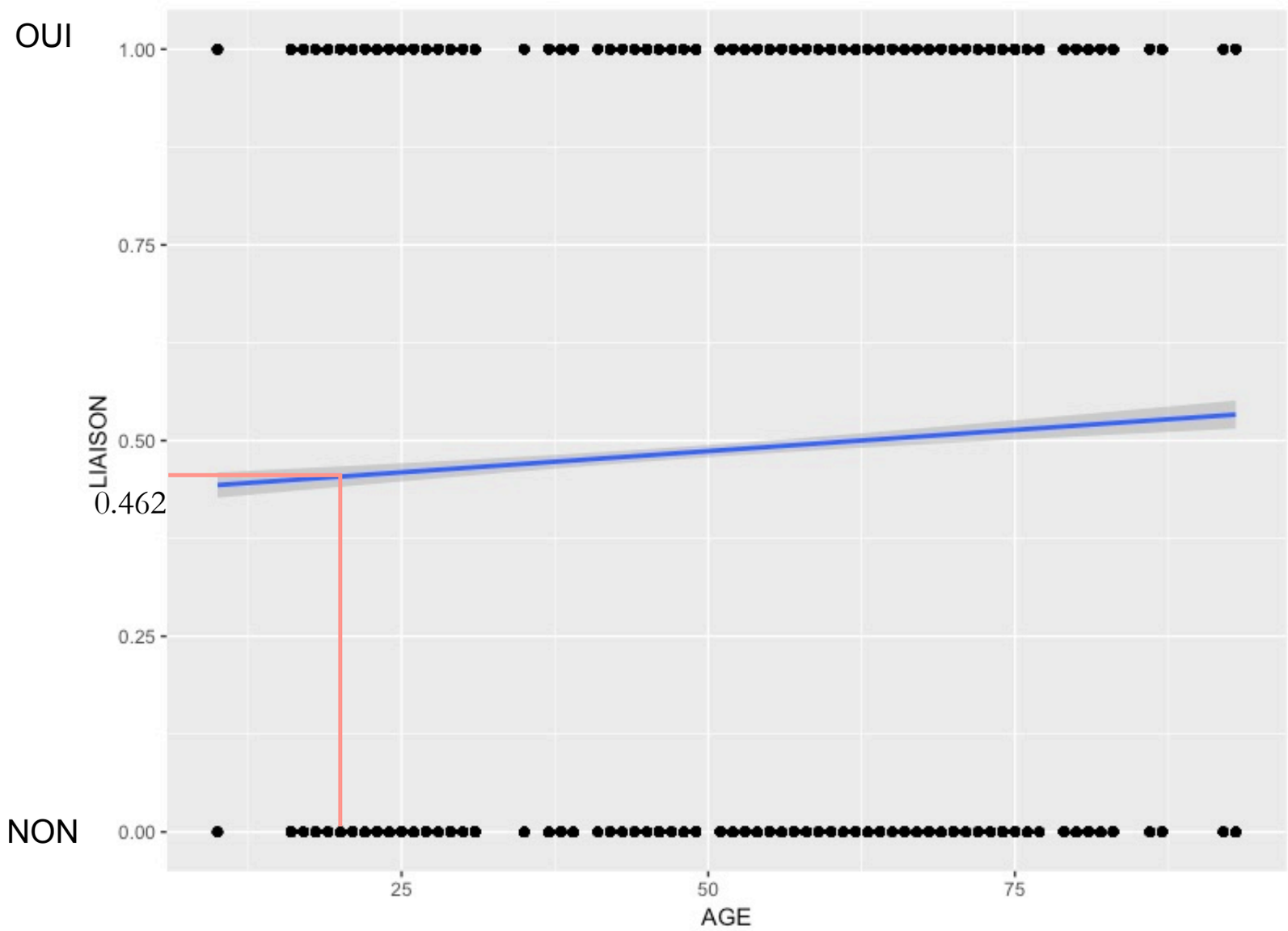
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



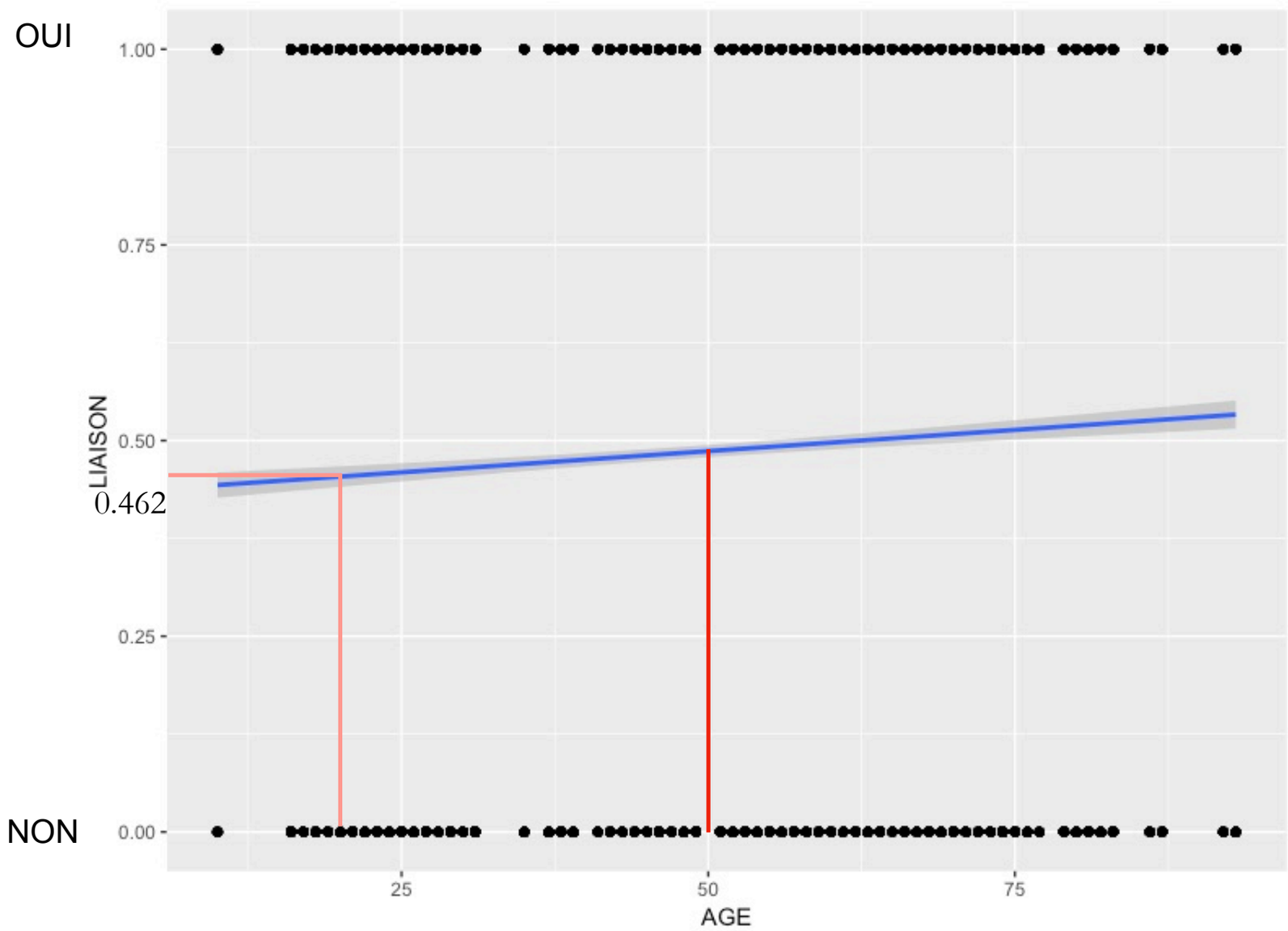
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



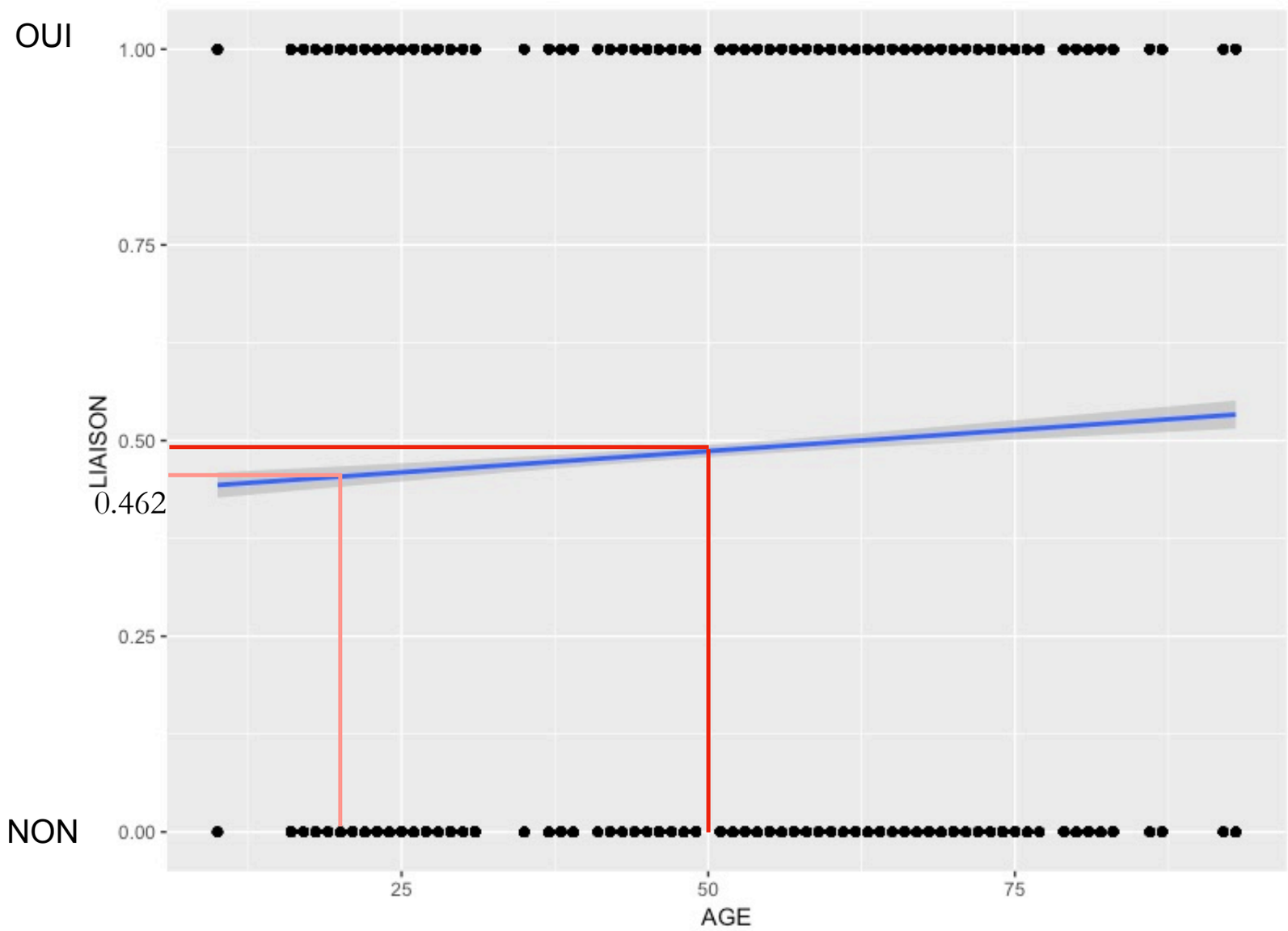
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



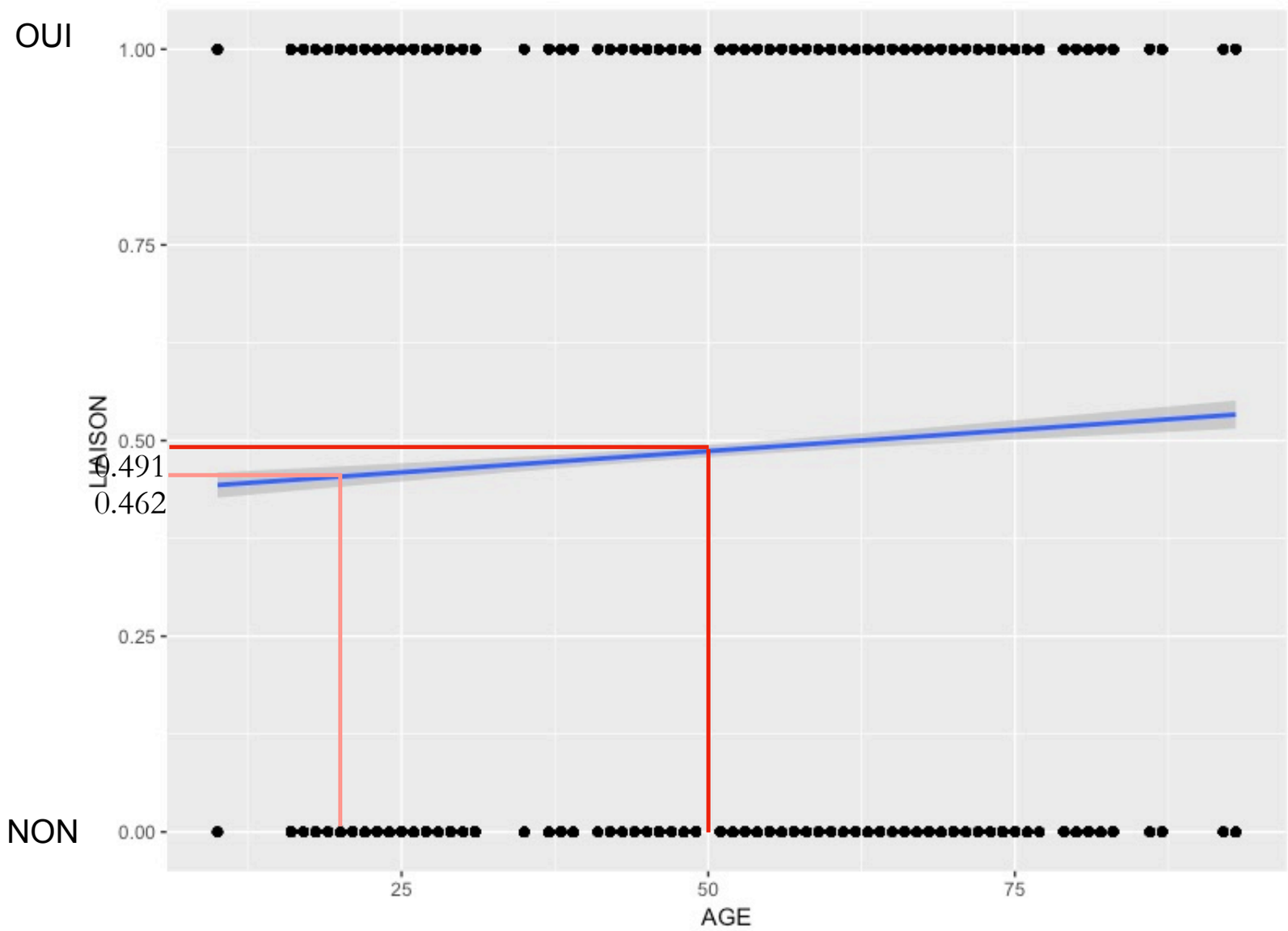
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



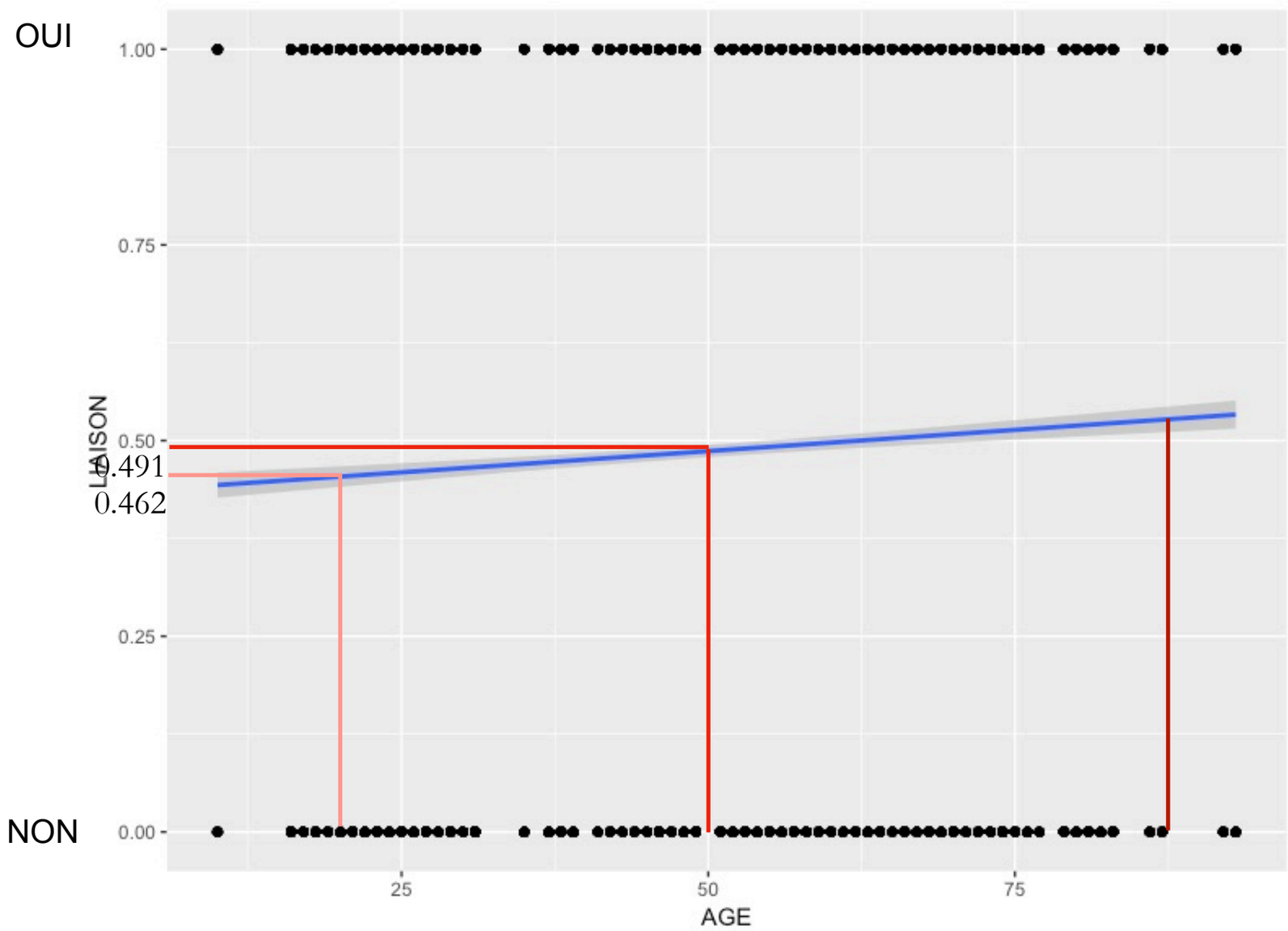
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



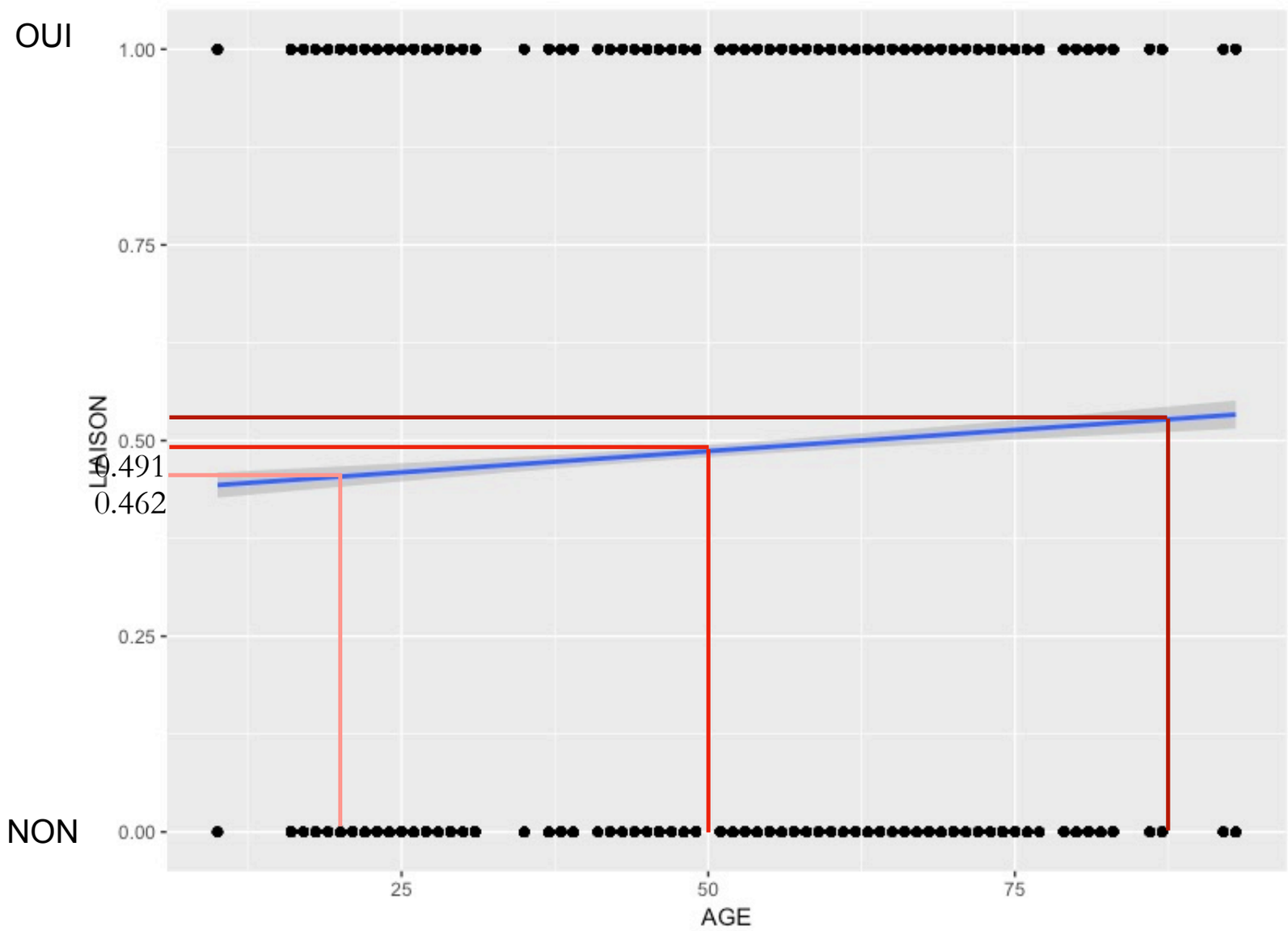
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



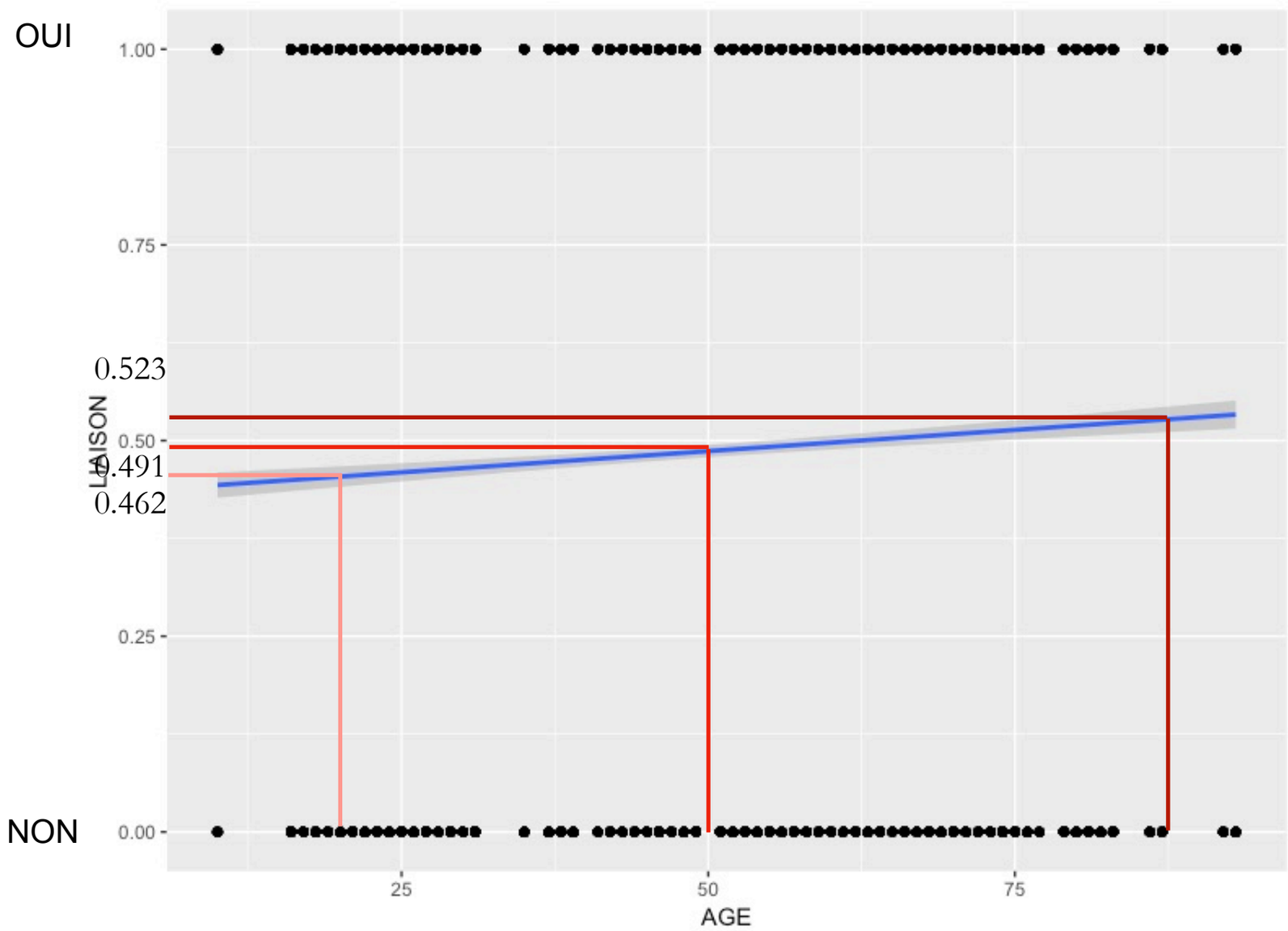
$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



$$P(Liaison = oui \mid AGE = 20) = 0.462$$
$$P(Liaison = oui \mid AGE = 50) = 0.491$$
$$P(Liaison = oui \mid AGE = 90) = 0.523$$



$$Y = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \epsilon$$

$$Y = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \epsilon$$

On essaie de trouver un relation entre l'AGE et le SEXE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 AGE + \beta_2 SEXE + \epsilon$$

$$Y = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \epsilon$$

On essaie de trouver un relation entre l'AGE et le SEXE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 AGE + \beta_2 SEXE + \epsilon$$

Coefficients:

(Intercept)	-0.2515769
AGE	0.0044065
SEXE-H	-0.0476284

$$Y = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \epsilon$$

On essaie de trouver un relation entre l'AGE et le SEXE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 AGE + \beta_2 SEXE + \epsilon$$

Coefficients:

(Intercept)	-0.2515769
AGE	0.0044065
SEXE-H	-0.0476284

$$Y = -0.2515769 + 0.0044065 \text{ AGE} + (-0.0476284) \text{ SEXE-H} + \epsilon$$

$$Y = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \epsilon$$

On essaie de trouver un relation entre l'AGE et le SEXE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 AGE + \beta_2 SEXE + \epsilon$$

Coefficients:

(Intercept)	-0.2515769
AGE	0.0044065
SEXE-H	-0.0476284

$$Y = -0.2515769 + 0.0044065\ \text{AGE} + (-0.0476284)\ \text{SEXE-H} + \epsilon$$

signe positif

+AGE +LIAISON (AGE *favorise* la réalisation)

variable quantitative

$$Y = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \epsilon$$

On essaie de trouver un relation entre l'AGE et le SEXE du locuteur et la réalisation de la liaison

$$Liaison = \beta_0 + \beta_1 AGE + \beta_2 SEXE + \epsilon$$

Coefficients:

(Intercept)	-0.2515769
AGE	0.0044065
SEXE-H	-0.0476284

$$Y = -0.2515769 + 0.0044065\ \text{AGE} + (-0.0476284)\ \text{SEXE-H} + \epsilon$$

signe positif
+AGE +LIAISON (AGE *favorise* la réalisation)

variable quantitative

signe négatif
-SEXE-H -LIAISON (SEXE-H *défavorise* la réalisation)

variable qualitative

Probabilité de réaliser une liaison par une femme de 70 ans

$$\mathrm{P}(\mathit{Liaison} = \mathit{oui} \mid \mathrm{AGE}{=}70,\ \mathrm{SEXE}{=}\mathrm{F}) = 0.4385201$$

Probabilité de réaliser une liaison par une femme de 70 ans

$$\mathrm{P}(Liaison = oui \mid \mathrm{AGE}{=}70,\ \mathrm{SEXE}{=}\mathrm{F}) = 0.4385201$$

Probabilité de réaliser une liaison par un homme de 70 ans

$$\mathrm{P}(Liaison = oui \mid \mathrm{AGE}{=}70,\ \mathrm{SEXE}{=}\mathrm{H}) = 0.4268295$$

Probabilité de réaliser une liaison par une femme de 70 ans

$$\mathrm{P}(Liaison = oui \mid \mathrm{AGE{=}70,\ SEXE{=}F}) = 0.4385201$$

Probabilité de réaliser une liaison par un homme de 70 ans

$$\mathrm{P}(Liaison = oui \mid \mathrm{AGE{=}70,\ SEXE{=}H}) = 0.4268295$$

Probabilité de réaliser une liaison par un homme de 18 ans

$$\mathrm{P}(Liaison = oui \mid \mathrm{AGE{=}18,\ SEXE{=}H}) = 0.3834355$$

Limites de la RL classique

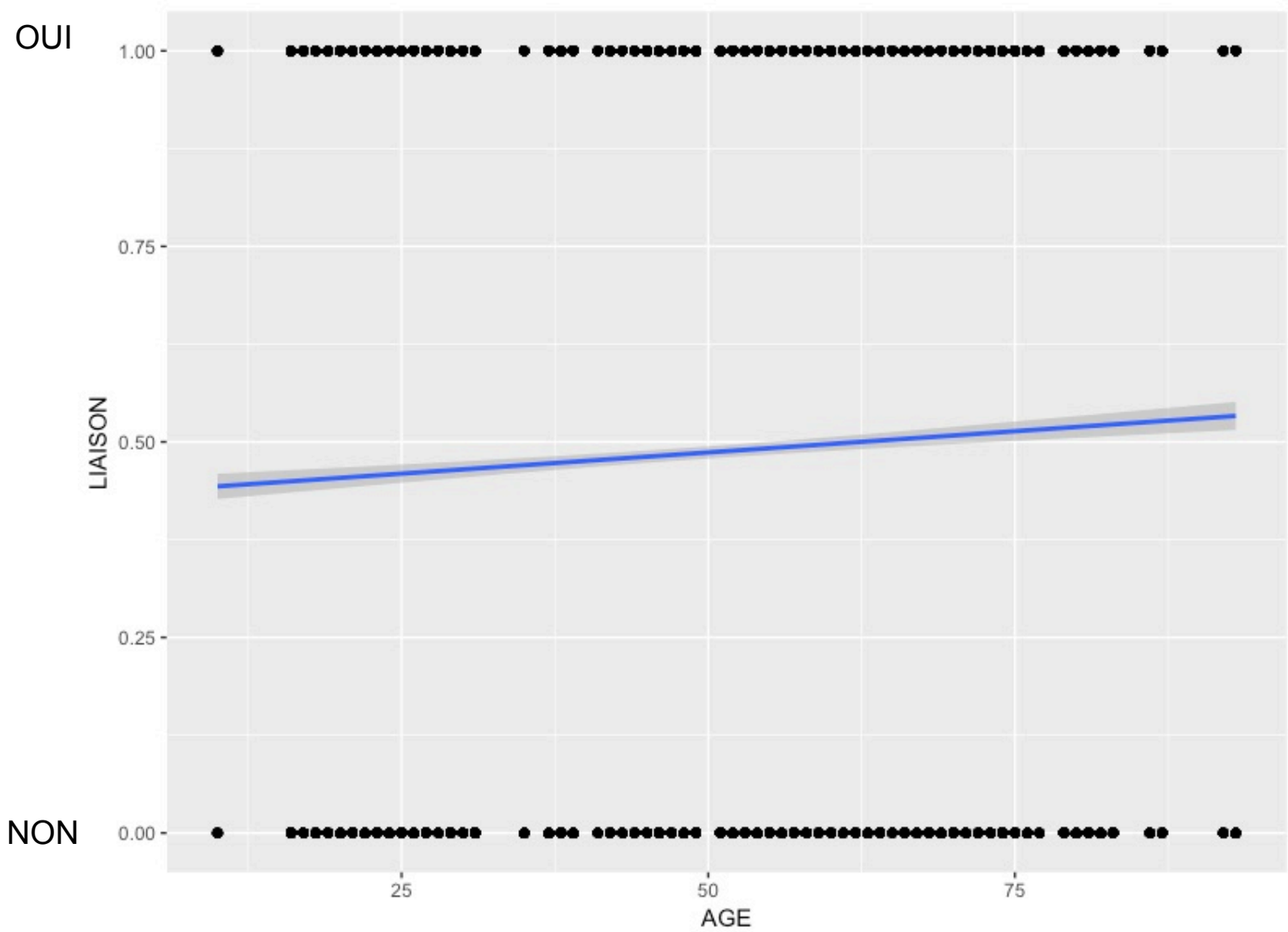
La RL se base sur des fonctions de relation entre Y (variable à expliquer) et X_n (variables explicatives) qui sont essentiellement linéaires

Limites de la RL classique

La RL se base sur des fonctions de relation entre Y (variable à expliquer) et X_n (variables explicatives) qui sont essentiellement linéaires

Possibilité d'utiliser des fonctions de lissage (*smooth functions*) qui ne sont pas forcement linéaires

Limites de la RL classique



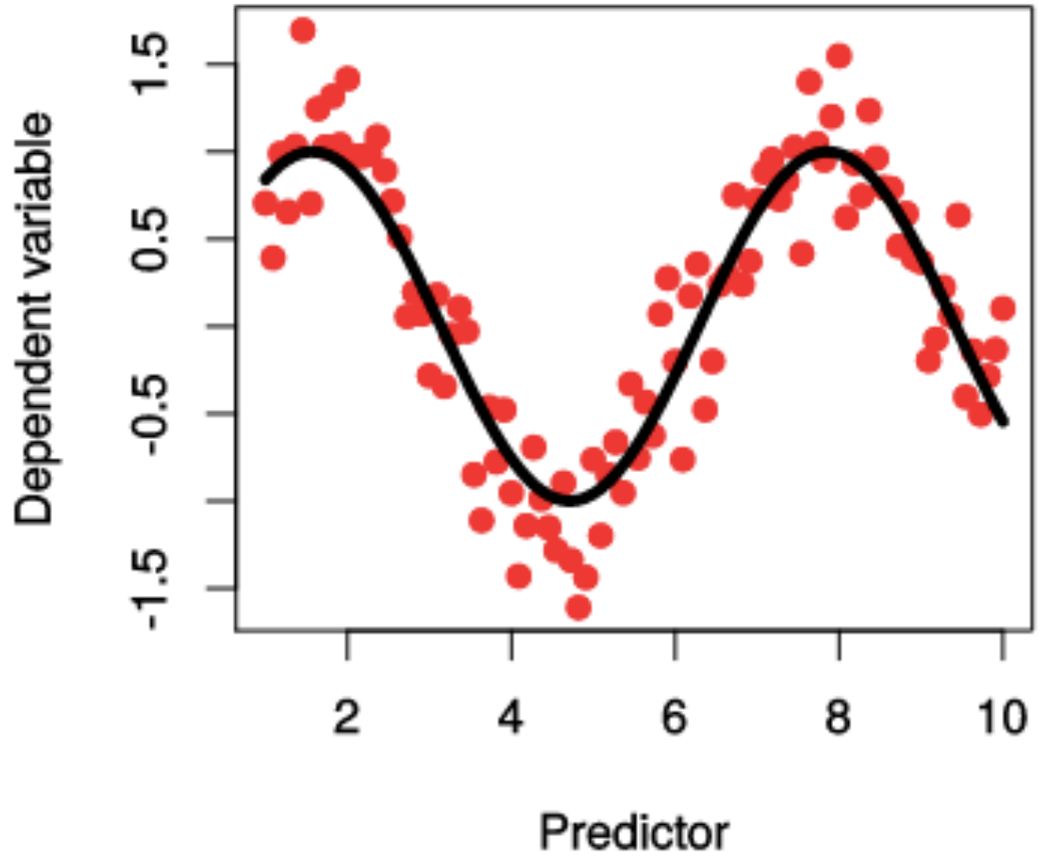
$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

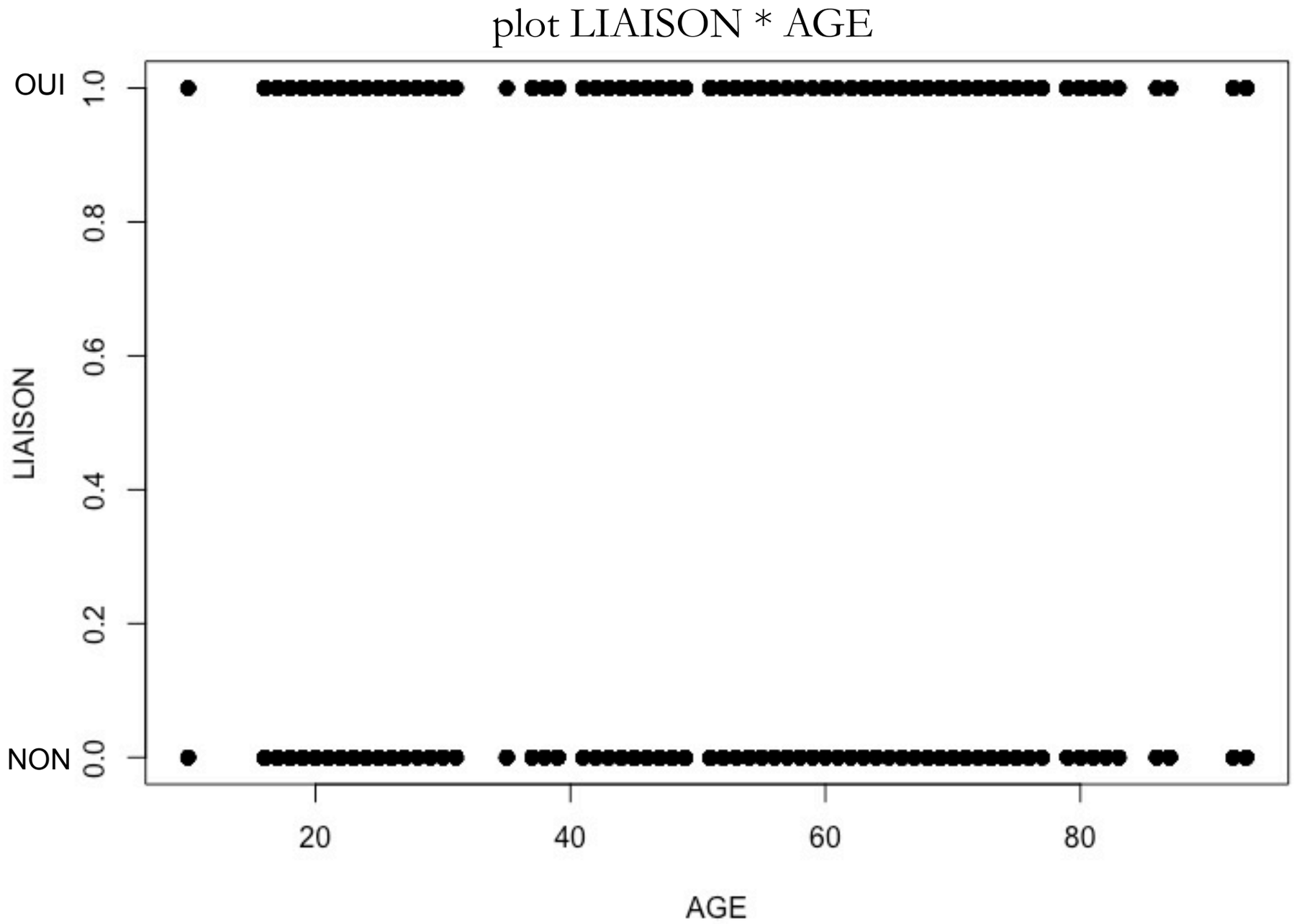
Les méthodes GAM s'appuient sur des transformations empiriques des variables explicatives par des techniques de *lissage*

$$Y = \beta_0 + \beta_1 \alpha_1 + \epsilon$$

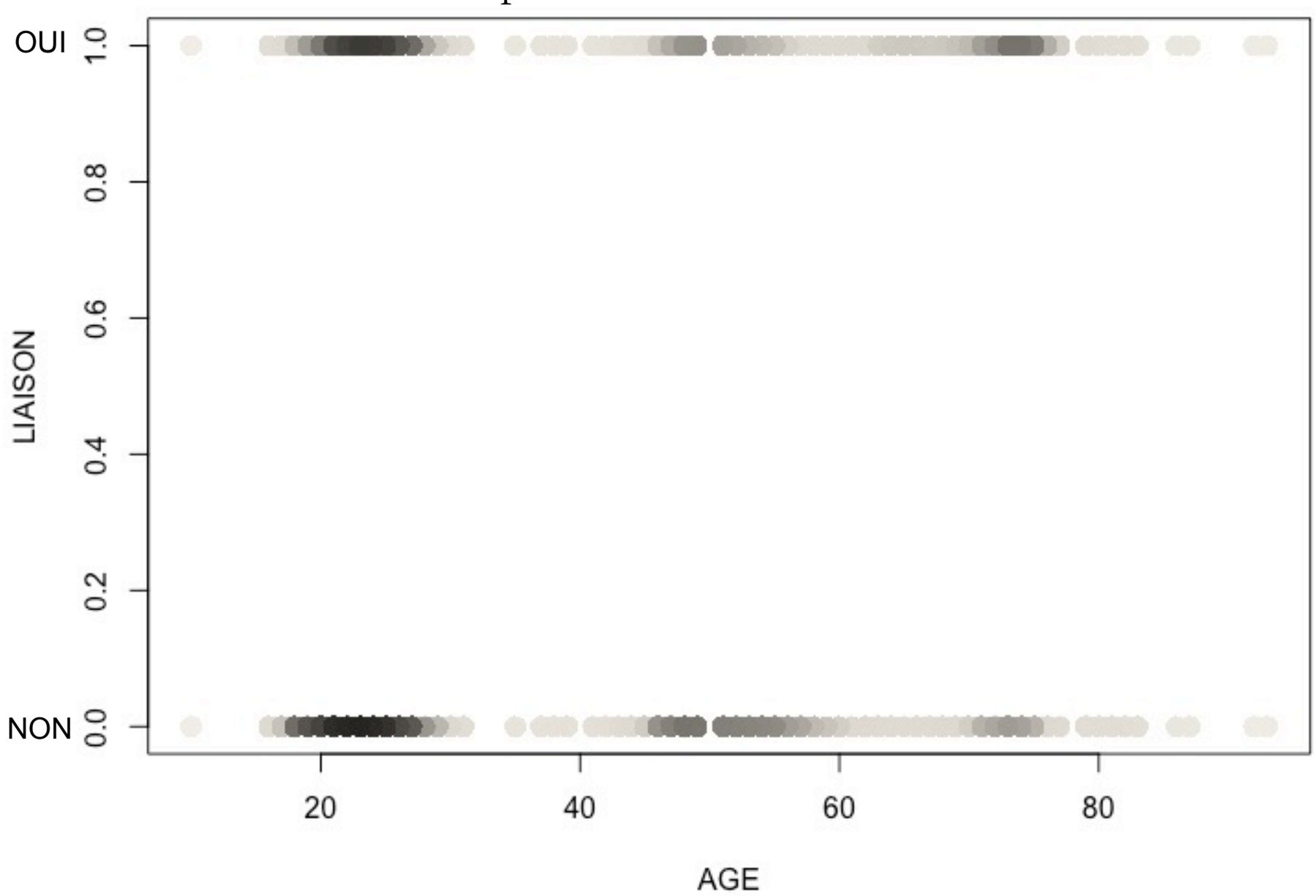
Les méthodes GAM s'appuient sur des transformations empiriques des variables explicatives par des techniques de *lissage*

$$g(E(Y)) = \beta_0 + f_1(\alpha_1) + \epsilon$$

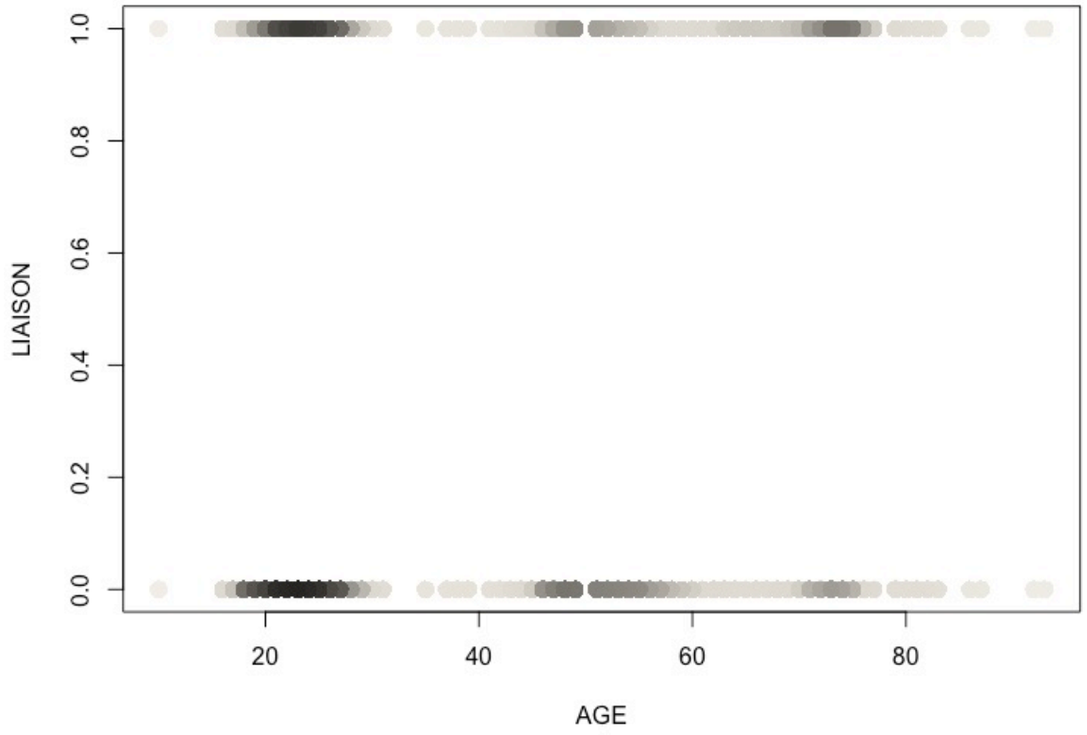



plot LIAISON * AGE
OUI
NON
LIAISON
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
AGE

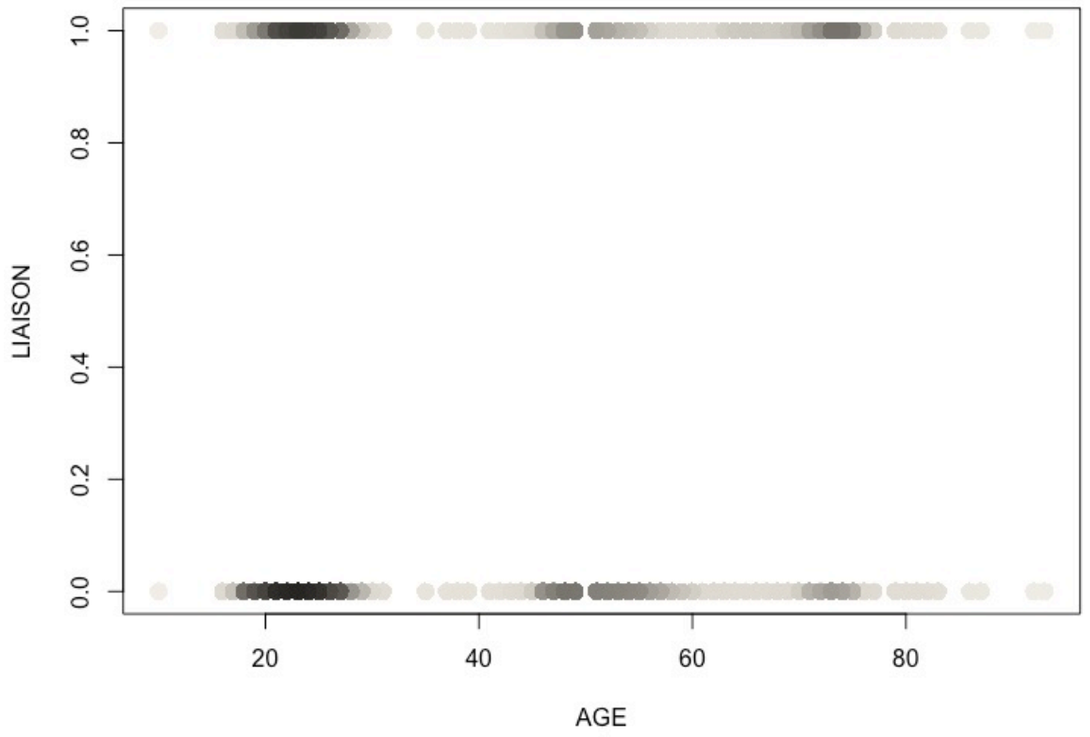
Visualisation de la densité



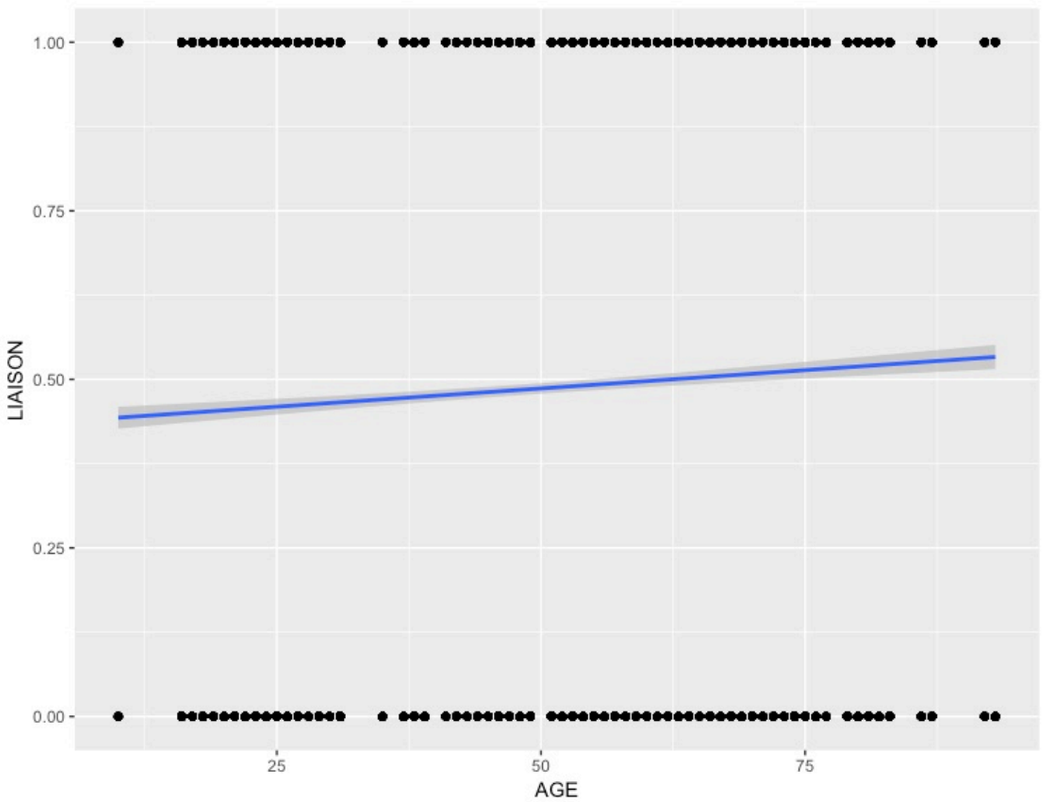
Fonction plus adaptative à la distribution réelle des données



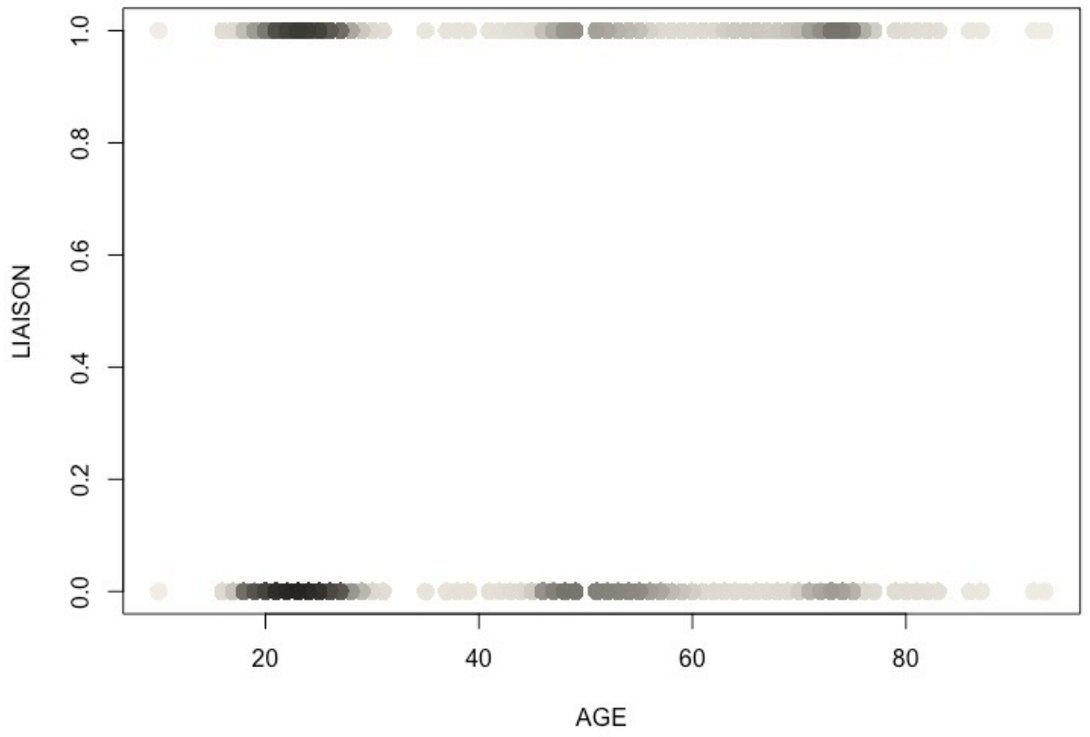
Fonction plus adaptative à la distribution réelle des données



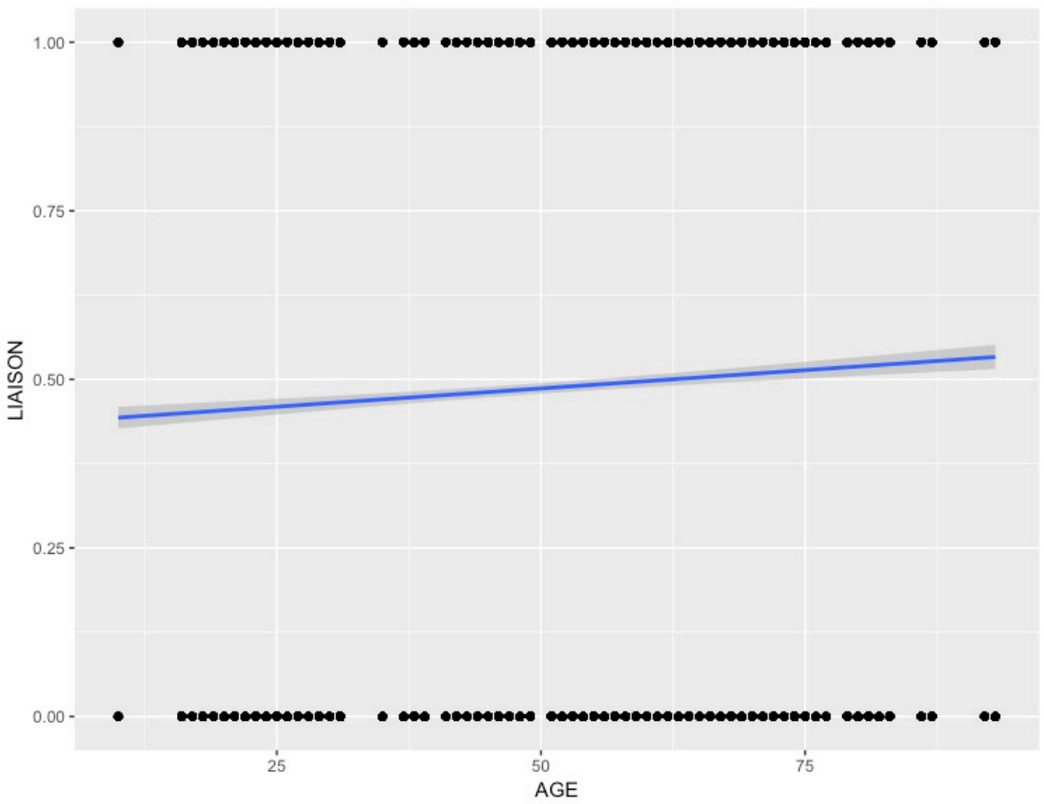
$$mod.AGE_{\mathrm{RL}} = \beta_0 + \beta_1 AGE$$



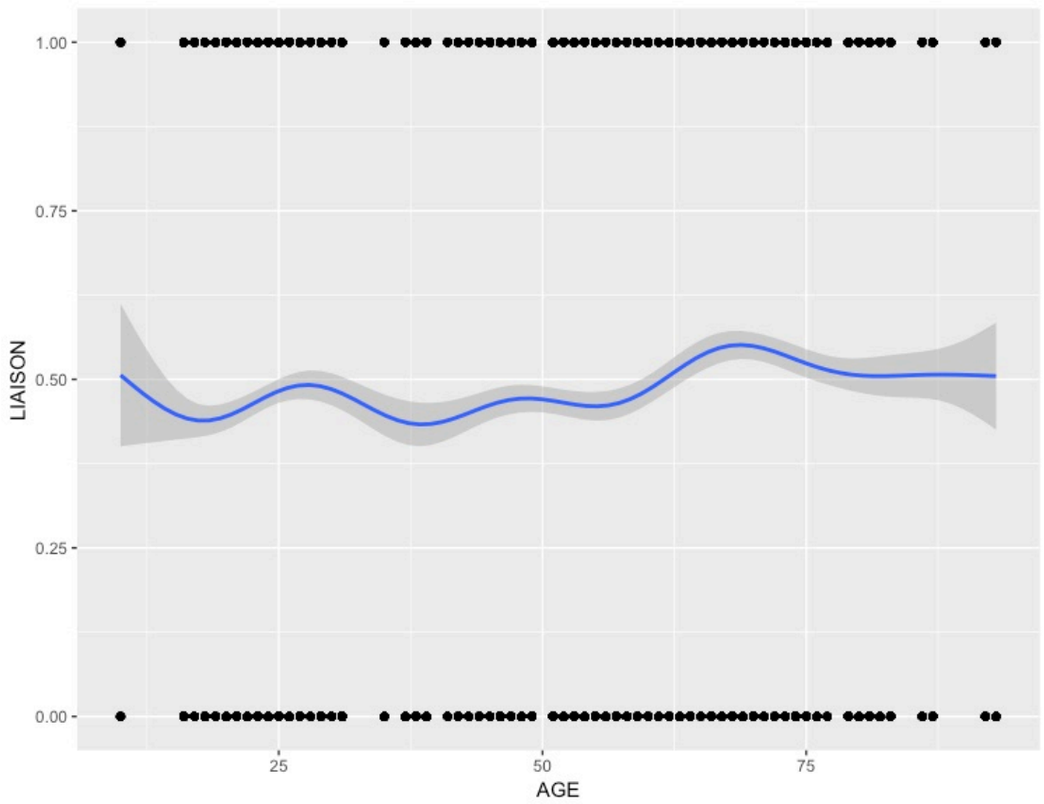
Fonction plus adaptative à la distribution réelle des données



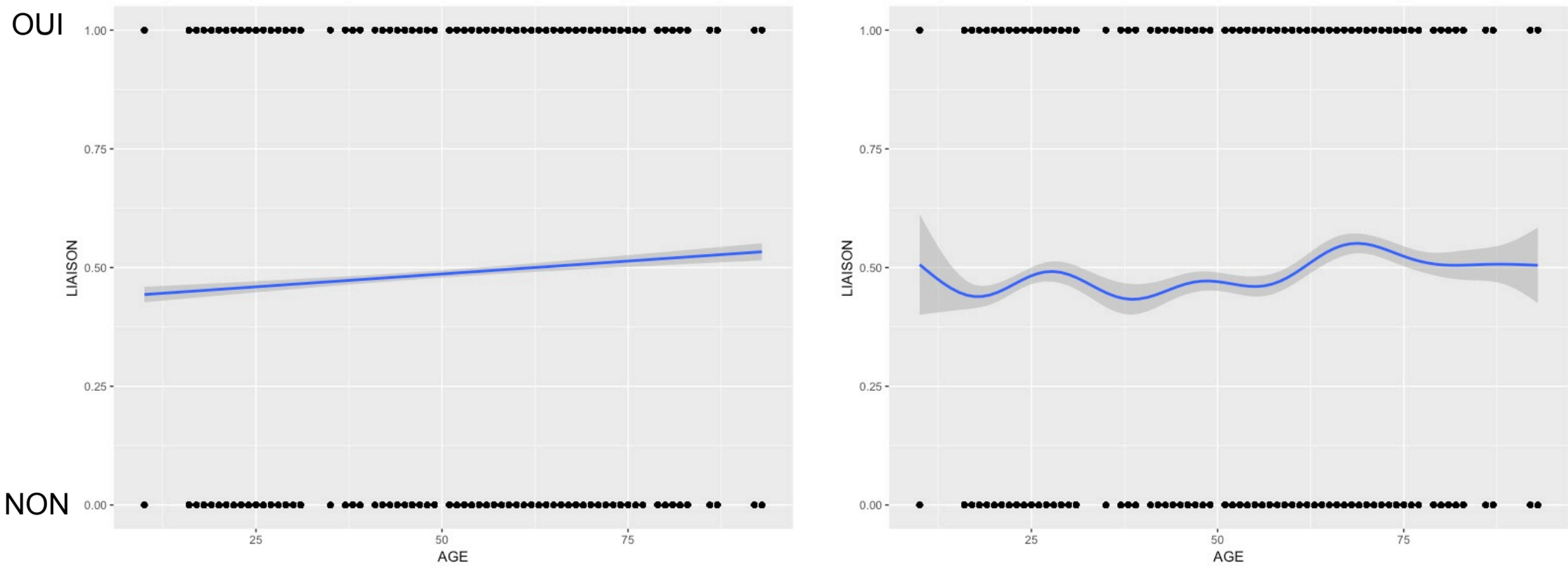
$$mod.AGE_{\mathrm{RL}} = \beta_0 + \beta_1 AGE$$



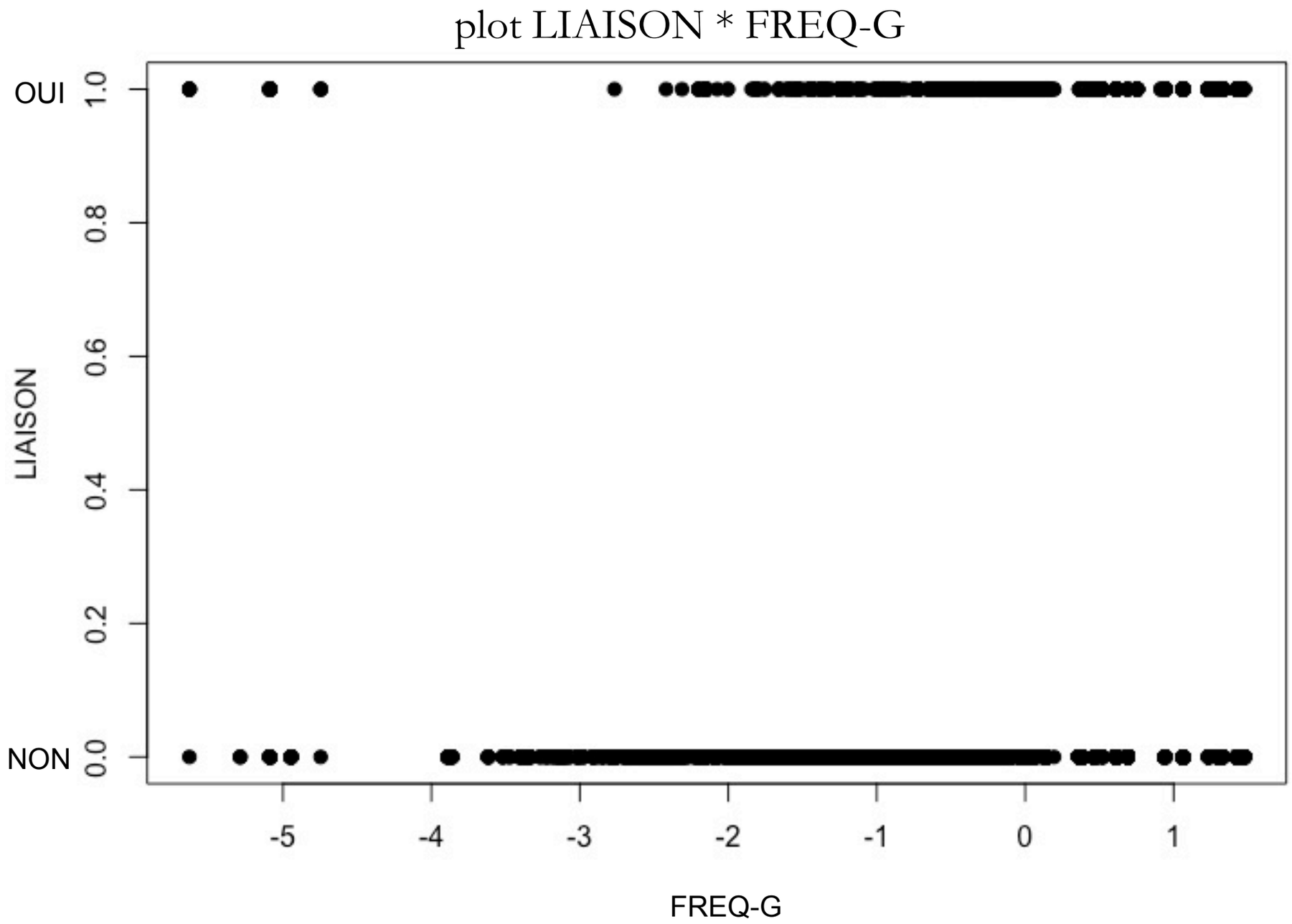
$$mod.AGE_{\mathrm{GAM}} = \beta_0 + s(AGE)$$



Comparaison des deux modèles

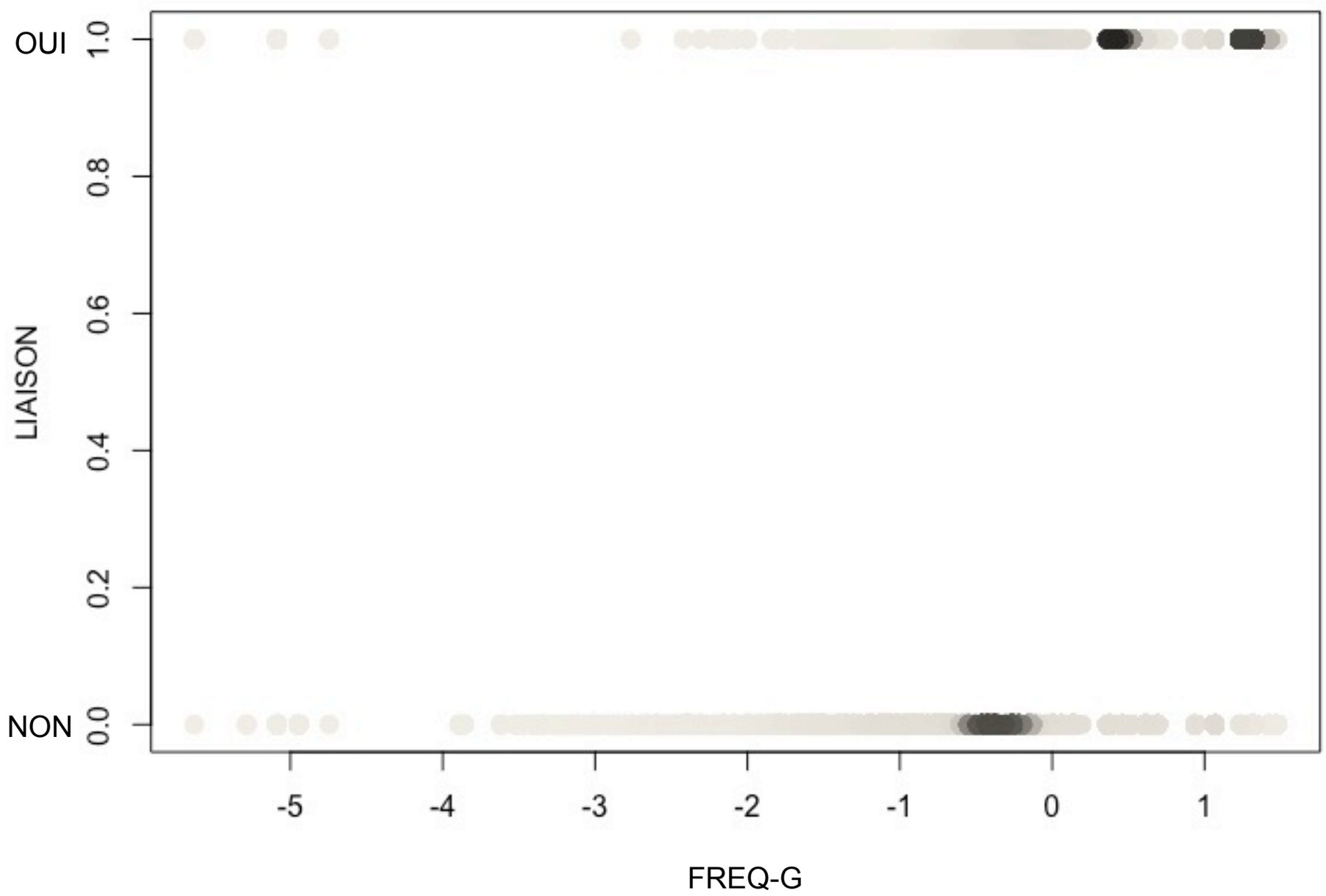


MODEL	AIC	%dev. expl.
mod.AGE$_{LR}$	21947.78	0.15
mod.AGE$_{GAM}$	21931.02	0.29

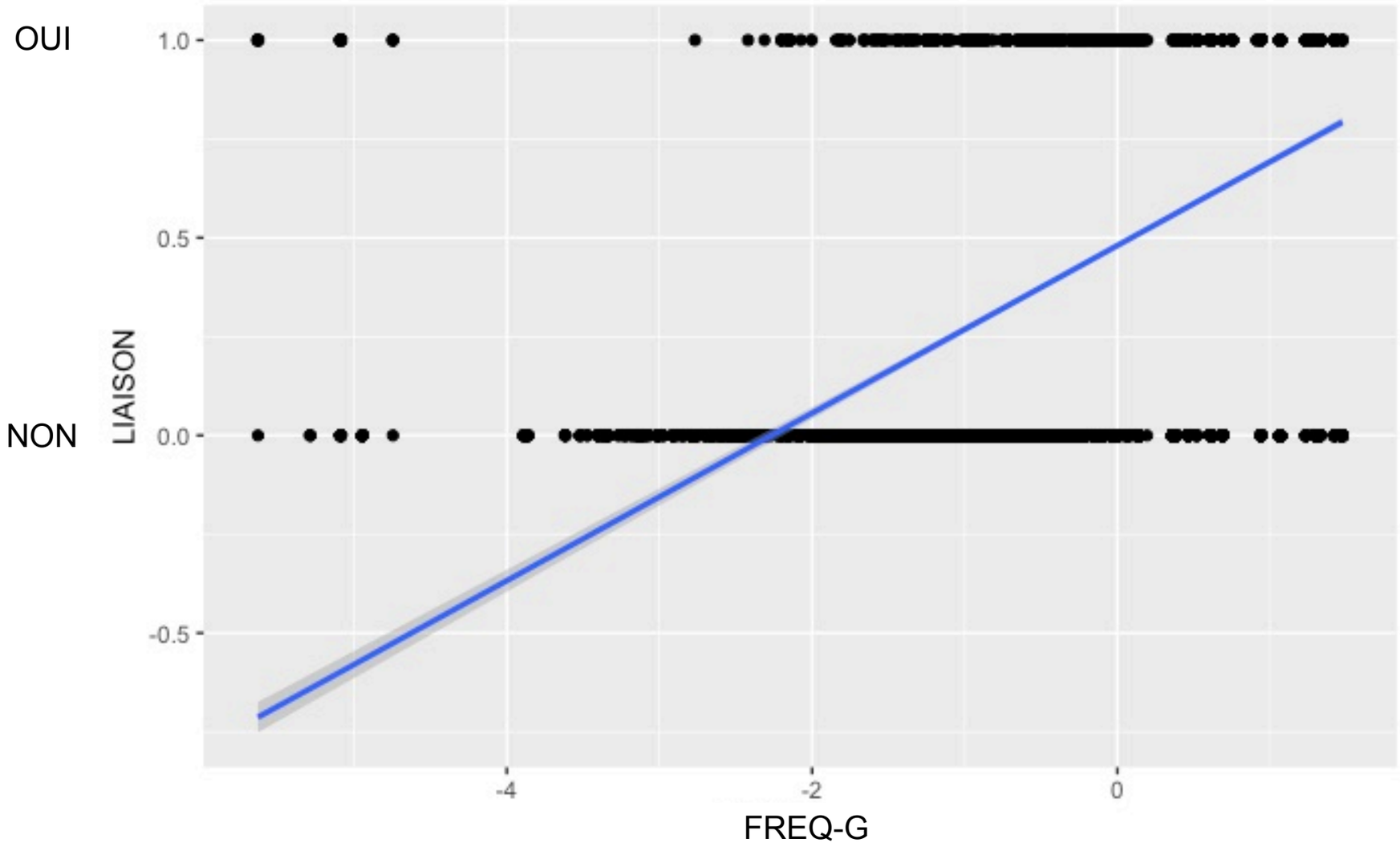

plot LIAISON * FREQ-G
OUI
NON
LIAISON
1.0
0.8
0.6
0.4
0.2
0.0
-5
-4
-3
-2
-1
0
1
FREQ-G

Visualisation de la densité

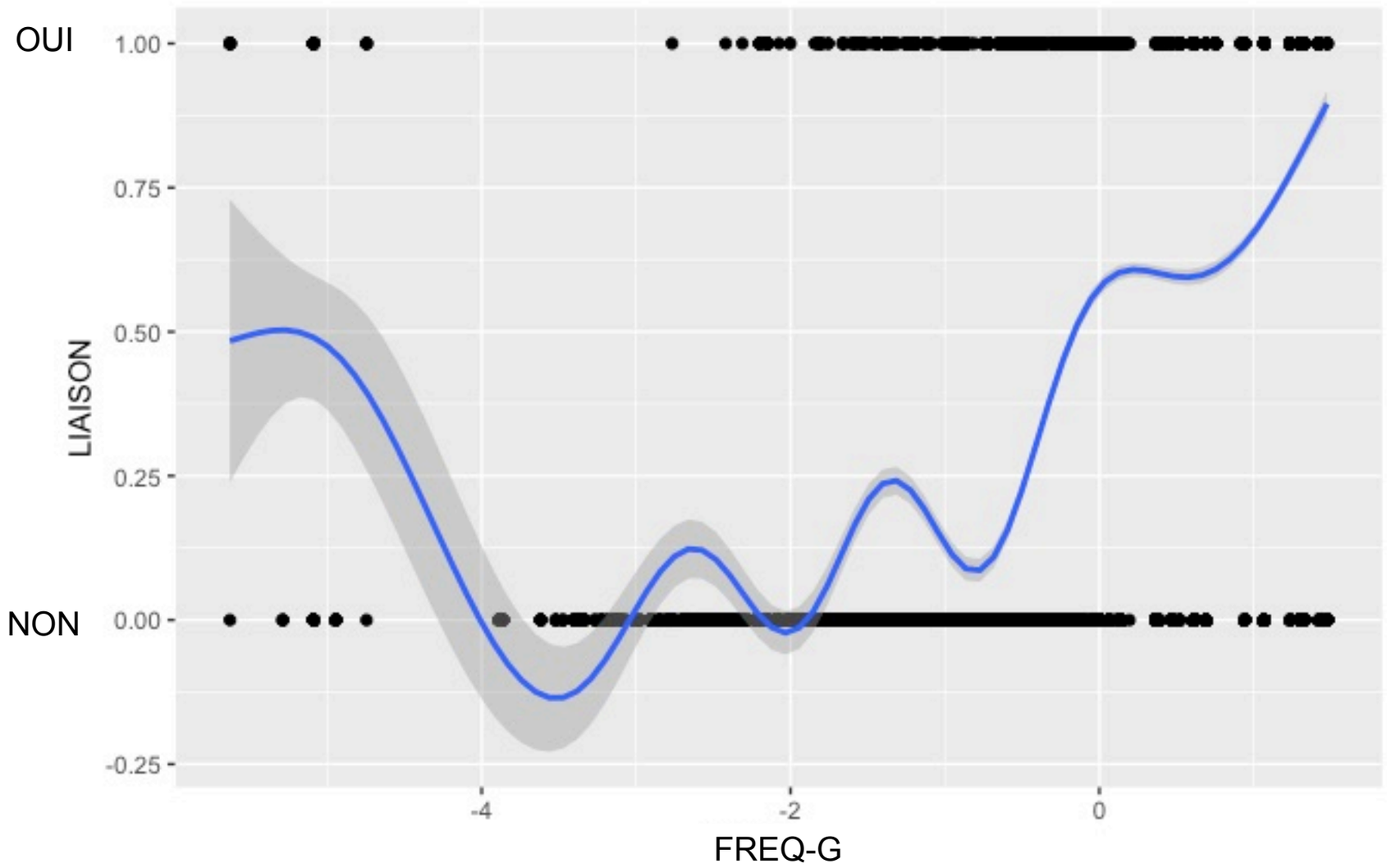
plot LIAISON * FREQ-G



$$mod.FREQ.G_{\mathrm{RL}} = \beta_0 + \beta_1 FREQ.G$$



$$mod.FREQ.G_{\text{GAM}} = \beta_0 + s(FREQ.G)$$



Une fois bien équipé avec ces outils statistiques on essaie de créer un model de prédiction de la liaison

Une fois bien équipé avec ces outils statistiques on essaie de créer un model de prédiction de la liaison

Calcul des variables outre que les variable présentes dans PFC, comme par exemple la fréquence des mots à gauche et à droite (information tirée par FrWaC) ou la longueur du mot phonologique (information tirée par la ressource PsychoGLAFF)

4.6 Les variables de notre modèle

LIAISON	LOCUTEUR	ÉDUCATION	SEXE	AGE	VILLE	MOT_GAUCHE	MOT_DROITE	POS_GAUCHE	POS_DROITE
OUI	21aml1	20	F	65	Dijon	temps	en	NOM	PRP
NON	44ajs2	14	F	59	Nantes	est	en	VER	PRP
…	…	…	…	…	…	…	…	…	…
oui = 8350 no = 9036 TOT = 17386	niveaux = 192	$\overline{\text{ÉDUCATION}}$ = 15,41	F = 9100 H = 8286	$\overline{\text{AGE}}$ = 48,11	niveaux = 20	niveaux = 487	niveaux = 487	niveaux = 11	niveaux = 7

L-SYLL_G	L-SYLL_D	L-PHONO_G	L-PHONO_D	FREQ_G	FREQ_D	F.BIGRAM	TP.BIGRAM	LATITUDE	LONGITUDE
1	2	1	3	4233	7655	433	0.09	47.32	5.04
2	3	2	2	190	651	115	0.13	47.21	1.55
…	…	…	…	…	…	…	…	…	…
$\overline{\text{SYLL\_G}}$ = 1,25	$\overline{\text{SYLL\_D}}$ = 1,84	$\overline{\text{PHONO\_G}}$ = 1,11	$\overline{\text{PHONO\_D}}$ = 2,31	$\overline{\text{FREQ\_G}}$ = 3213	$\overline{\text{FREQ\_D}}$ = 4390	$\overline{\text{F.BIGRAM}}$ = 3691	$\overline{\text{TP.BIGRAM}}$ = 0.02		

4.6 Les variables de notre modèle

LIAISON	LOCUTEUR	ÉDUCATION	SEXE	AGE	VILLE	MOT_GAUCHE	MOT_DROITE	POS_GAUCHE	POS_DROITE
OUI	21aml1	20	F	65	Dijon	temps	en	NOM	PRP
NON	44ajs2	14	F	59	Nantes	est	en	VER	PRP
…	…	…	…	…	…	…	…	…	…
oui = 8350 no = 9036 TOT = 17386	niveaux = 192	$\overline{\text{ÉDUCATION}}$ = 15,41	F = 9100 H = 8286	$\overline{\text{AGE}}$ = 48,11	niveaux = 20	niveaux = 487	niveaux = 487	niveaux = 11	niveaux = 7

Toutes les variables quantitatives sont centrées réduites (c'est-à-dire : moyenne = 0 et écart-type = 1)

L-SYLL_G	L-SYLL_D	L-PHONO_G	L-PHONO_D	FREQ_G	FREQ_D	F.BIGRAM	TP.BIGRAM	LATITUDE	LONGITUDE
1	2	1	3	4233	7655	433	0.09	47.32	5.04
2	3	2	2	190	651	115	0.13	47.21	1.55
…	…	…	…	…	…	…	…	…	…
$\overline{\text{SYLL\_G}}$ = 1,25	$\overline{\text{SYLL\_D}}$ = 1,84	$\overline{\text{PHONO\_G}}$ = 1,11	$\overline{\text{PHONO\_D}}$ = 2,31	$\overline{\text{FREQ\_G}}$ = 3213	$\overline{\text{FREQ\_D}}$ = 4390	$\overline{\text{F.BIGRAM}}$ = 3691	$\overline{\text{TP.BIGRAM}}$ = 0.02		

4.6 Les variables de notre modèle

LIAISON	LOCUTEUR	ÉDUCATION	SEXE	AGE	VILLE	MOT_GAUCHE	MOT_DROITE	POS_GAUCHE	POS_DROITE
OUI	21aml1	20	F	65	Dijon	temps	en	NOM	PRP
NON	44ajs2	14	F	59	Nantes	est	en	VER	PRP
…	…	…	…	…	…	…	…	…	…
oui = 8350 no = 9036 TOT = 17386	niveaux = 192	$\overline{\text{ÉDUCATION}}$ = 15,41	F = 9100 H = 8286	$\overline{\text{AGE}}$ = 48,11	niveaux = 20	niveaux = 487	niveaux = 487	niveaux = 11	niveaux = 7

Toutes les variables quantitatives sont centrées réduites (c'est-à-dire : moyenne = 0 et écart-type = 1)

Toutes les variables *FREQ* sont centrées réduites en fonction *log*

L-SYLL_G	L-SYLL_D	L-PHONO_G	L-PHONO_D	FREQ_G	FREQ_D	F.BIGRAM	TP.BIGRAM	LATITUDE	LONGITUDE
1	2	1	3	4233	7655	433	0.09	47.32	5.04
2	3	2	2	190	651	115	0.13	47.21	1.55
…	…	…	…	…	…	…	…	…	…
$\overline{\text{SYLL\_G}}$ = 1,25	$\overline{\text{SYLL\_D}}$ = 1,84	$\overline{\text{PHONO\_G}}$ = 1,11	$\overline{\text{PHONO\_D}}$ = 2,31	$\overline{\text{FREQ\_G}}$ = 3213	$\overline{\text{FREQ\_D}}$ = 4390	$\overline{\text{F.BIGRAM}}$ = 3691	$\overline{\text{TP.BIGRAM}}$ = 0.02		

La méthode adoptée pour la sélection des variables définissant le modèle est la régression itérative (procédure *stepwise progressive*) qui inclut d'abord la variable qui propose la meilleur déviance expliquée.
Ensuite, celle qui améliore le plus la déviance et ainsi de suite.

La méthode adoptée pour la sélection des variables définissant le modèle est la régression itérative (procédure *stepwise progressive*) qui inclut d'abord la variable qui propose la meilleur déviance expliquée.
Ensuite, celle qui améliore le plus la déviance et ainsi de suite.

$$model1 = \beta_0 + s(MOT - G)$$

La méthode adoptée pour la sélection des variables définissant le modèle est la régression itérative (procédure *stepwise progressive*) qui inclut d'abord la variable qui propose la meilleur déviance expliquée.
Ensuite, celle qui améliore le plus la déviance et ainsi de suite.

$$model1 = \beta_0 + s(MOT - G)$$
$$model2 = \beta_0 + s(MOT - G) + s(FREQ - G)$$

La méthode adoptée pour la sélection des variables définissant le modèle est la régression itérative (procédure *stepwise progressive*) qui inclut d'abord la variable qui propose la meilleur déviance expliquée.
Ensuite, celle qui améliore le plus la déviance et ainsi de suite.

Création des modèles : calcul de AIC et (%) déviance expliquée

$$model1 = \beta_0 + s(MOT - G)$$
$$model2 = \beta_0 + s(MOT - G) + s(FREQ - G)$$

La méthode adoptée pour la sélection des variables définissant le modèle est la régression itérative (procédure *stepwise progressive*) qui inclut d'abord la variable qui propose la meilleur déviance expliquée.
Ensuite, celle qui améliore le plus la déviance et ainsi de suite.

Création des modèles : calcul de AIC et (%) déviance expliquée

$$
\begin{aligned}
model1 &= \beta_0 + s(MOT - G) \\
model2 &= \beta_0 + s(MOT - G) + s(FREQ - G) \\
model3 &= \beta_0 + s(MOT - G) + s(FREQ - G) + s(POS - D) \\
model4 &= \beta_0 + s(MOT - G) + s(FREQ - G) + s(POS - D) + s(TP.BIGRAM) \\
&\dots
\end{aligned}
$$

Predictors	**Type**	**%dev. expl.**	**AIC**	**t-test**
locuteur	factor (192)	3.72	23479.67	$p < 0.001 * **$
age	numer.	0.29	21931.02	$p < 0.001 * **$
éducation	numer.	0.63	15388.94	$p < 0.001 * **$
sexe	factor (2)	0.02	24074.83	$p < 0.001 * **$
mot-gauche	factor (487)	**70.80**	**7699.546**	$p < 0.001 * **$
mot-droite	factor (1531)	**48.41**	**14551.51**	$p < 0.001 * **$
pos-gauche	factor (11)	**49.91**	**11933.72**	$p < 0.001 * **$
pos-droite	factor (7)	**23.00**	**17260.84**	$p < 0.001 * **$
long.syll-g	numer.	16.50	20081.17	$p < 0.001 * **$
long.syll-d	numer.	9.30	22932.11	$p < 0.001 * **$
long.phono-g	numer.	15.43	21341.09	$p < 0.001 * **$
long.phono-d	numer.	7.52	22231.40	$p < 0.001 * **$
freq-g	numer.	**18.41**	**19662.86**	$p < 0.001 * **$
freq-d	numer.	1.32	26731.56	$p < 0.001 * **$
f.bigram	numer.	7.97	20633.03	$p < 0.001 * **$
tp.bigram	numer.	5.20	21252.77	$p < 0.001 * **$
(longitude,latitude)	numer.	1.72	23923.42	$p < 0.001 * **$

Predictors	**Type**	**%dev. expl.**	**AIC**	**t-test**
locuteur	factor (192)	3.72	23479.67	$p < 0.001 * **$
age	numer.	0.29	21931.02	$p < 0.001 * **$
éducation	numer.	0.63	15388.94	$p < 0.001 * **$
sexe	factor (2)	0.02	24074.83	$p < 0.001 * **$
mot-gauche	factor (487)	**70.80**	**7699.546**	$p < 0.001 * **$
mot-droite	factor (1531)	**48.41**	**14551.51**	$p < 0.001 * **$
pos-gauche	factor (11)	**49.91**	**11933.72**	$p < 0.001 * **$
pos-droite	factor (7)	**23.00**	**17260.84**	$p < 0.001 * **$
long.syll-g	numer.	16.50	20081.17	$p < 0.001 * **$
long.syll-d	numer.	9.30	22932.11	$p < 0.001 * **$
long.phono-g	numer.	15.43	21341.09	$p < 0.001 * **$
long.phono-d	numer.	7.52	22231.40	$p < 0.001 * **$
freq-g	numer.	**18.41**	**19662.86**	$p < 0.001 * **$
freq-d	numer.	1.32	26731.56	$p < 0.001 * **$
f.bigram	numer.	7.97	20633.03	$p < 0.001 * **$
tp.bigram	numer.	5.20	21252.77	$p < 0.001 * **$
(longitude,latitude)	numer.	1.72	23923.42	$p < 0.001 * **$

- Objectif du travail
- La liaison - une petite présentation
- Un corpus phonologique : PFC
- Les outils statistiques (RL et GAM)
- Le model final et le résultats
- Limites du travail et perspectives futures

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ{-}G) + s(SYLL{-}G) + s(TP.BIGRAM) + s(MOT{-}G) +$$

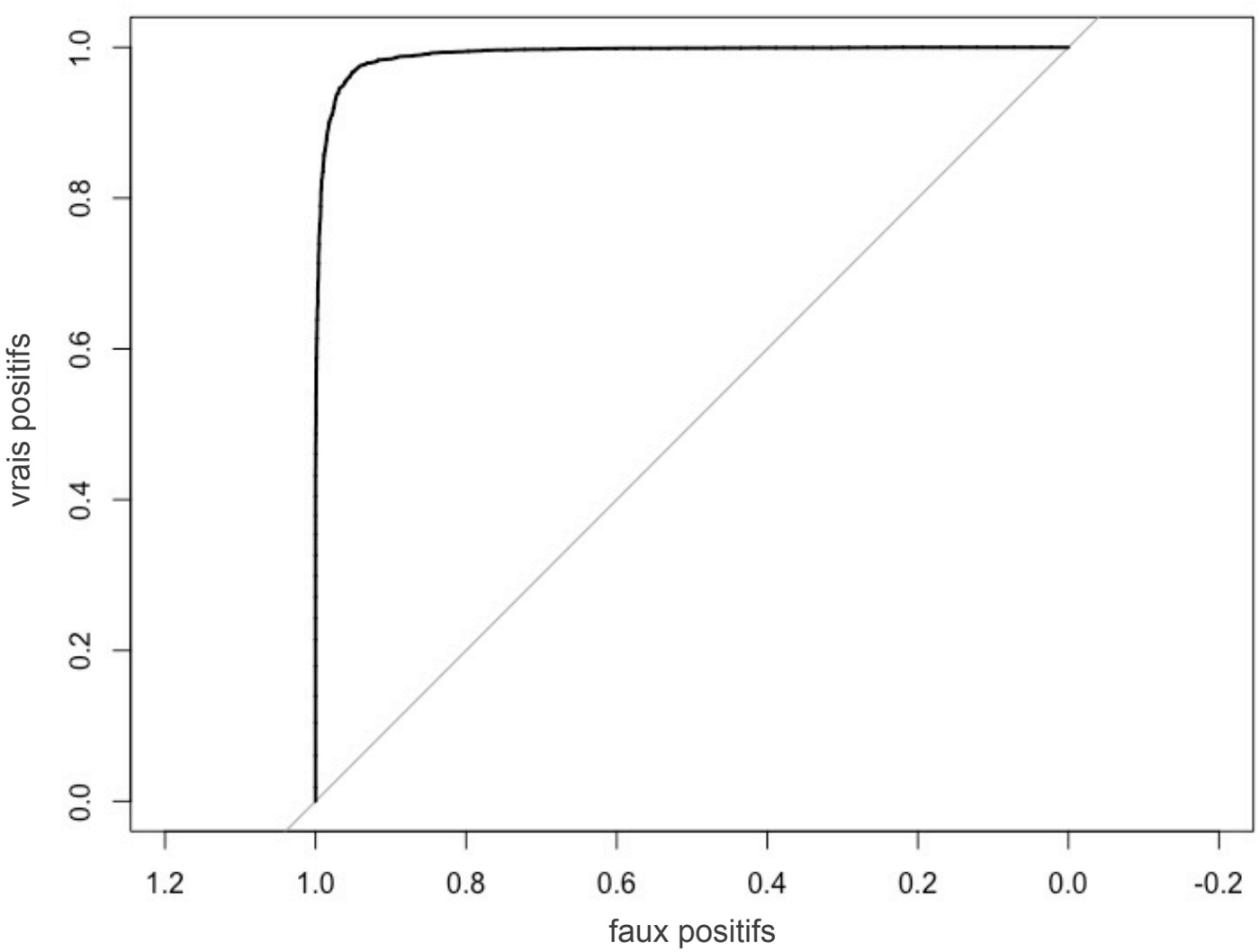
$$+ s(POS{-}G) + s(POS{-}D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

AIC	2999.090
% dev. expl.	83.214
AUC	0.991

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

AIC	2999.090
% dev. expl.	83.214
AUC	0.991



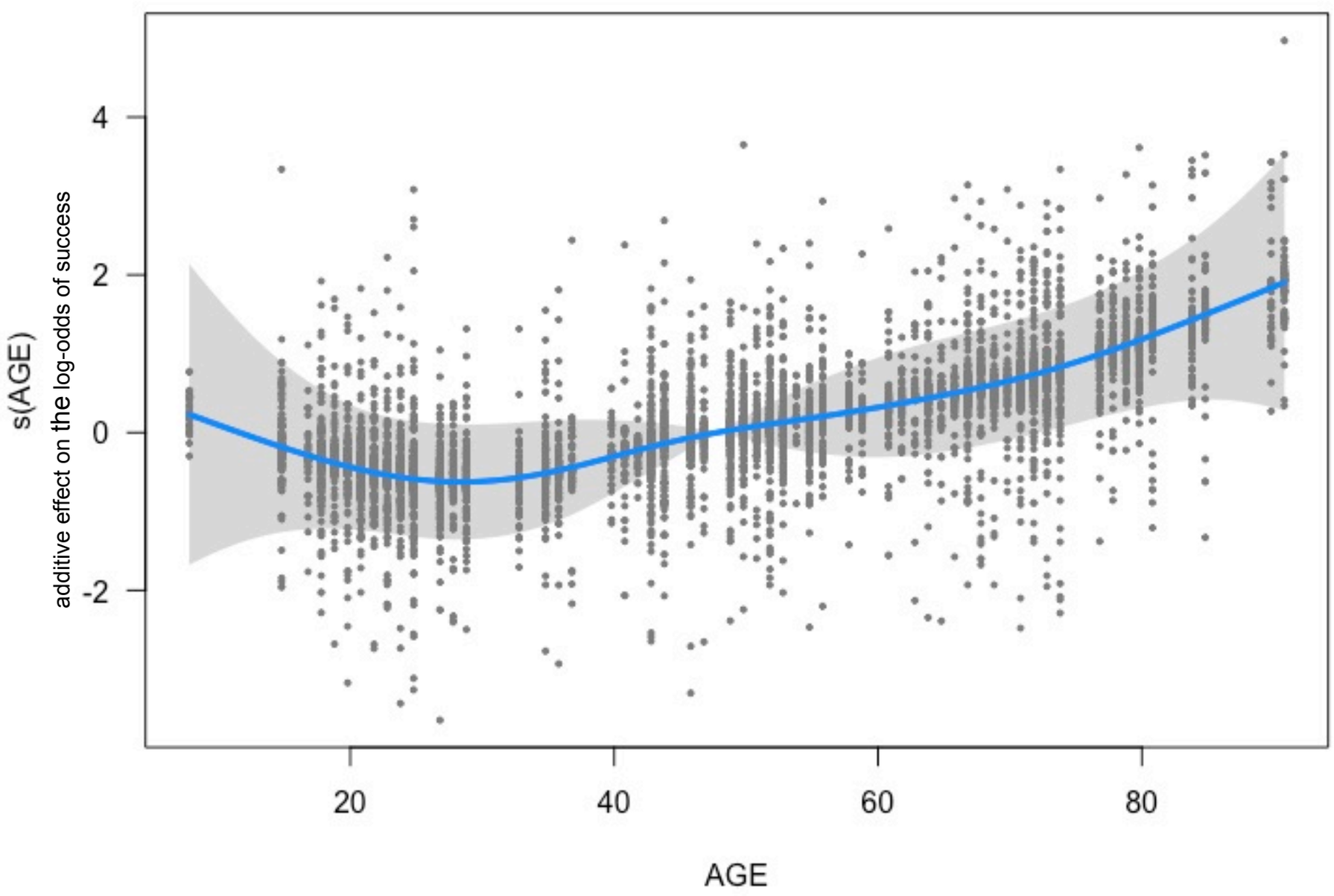
$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

6.1 Le modèle final - Les fonctions de lissage

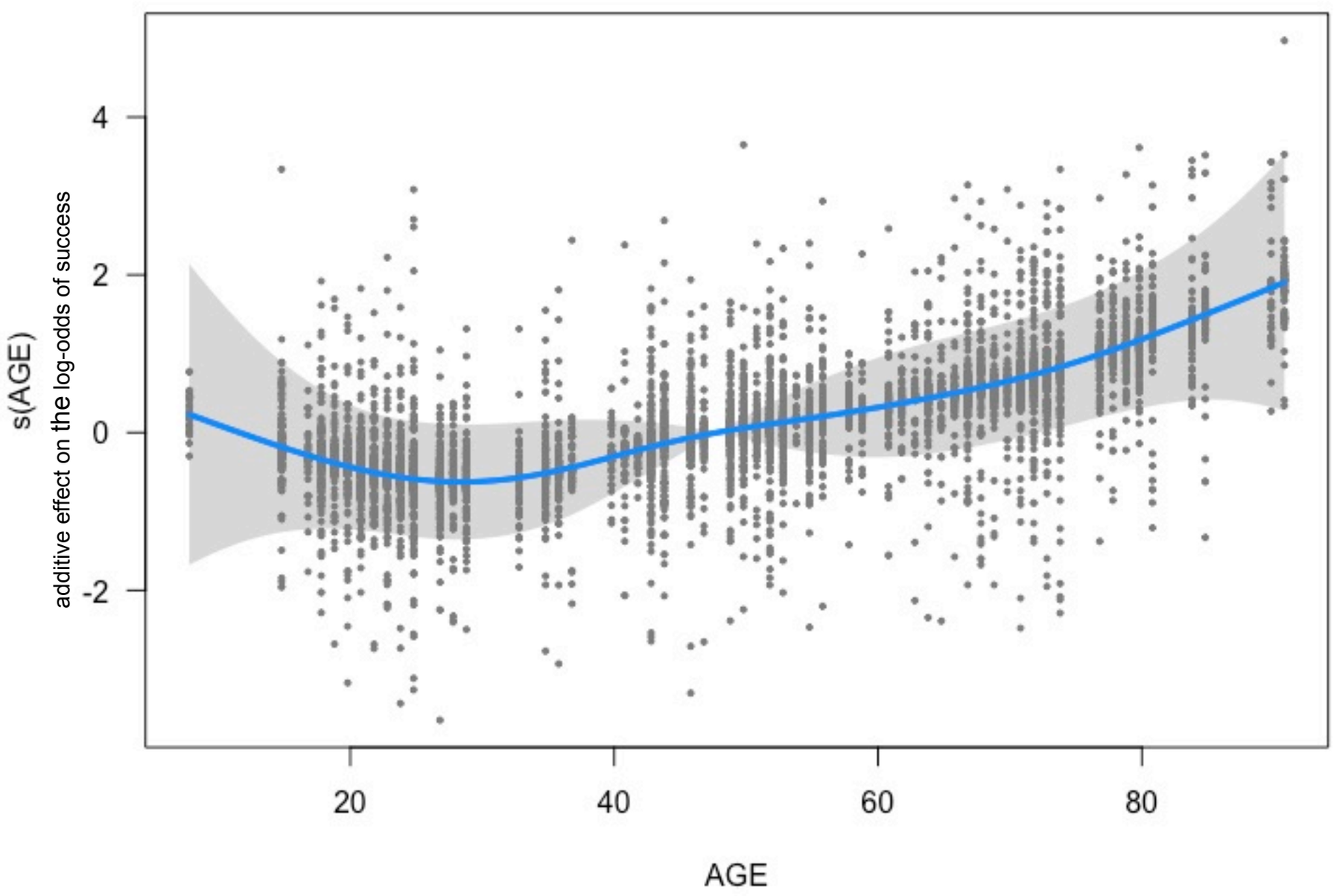
$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ\text{-}G) + s(SYLL\text{-}G) + s(TP.BIGRAM) + s(MOT\text{-}G) +$$
$$+ s(POS\text{-}G) + s(POS\text{-}D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ{-}G) + s(SYLL{-}G) + s(TP.BIGRAM) + s(MOT{-}G) +$$
$$+ s(POS{-}G) + s(POS{-}D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

L'augmentation de l'AGE détermine une plus forte probabilité d'avoir une liaison réalisée



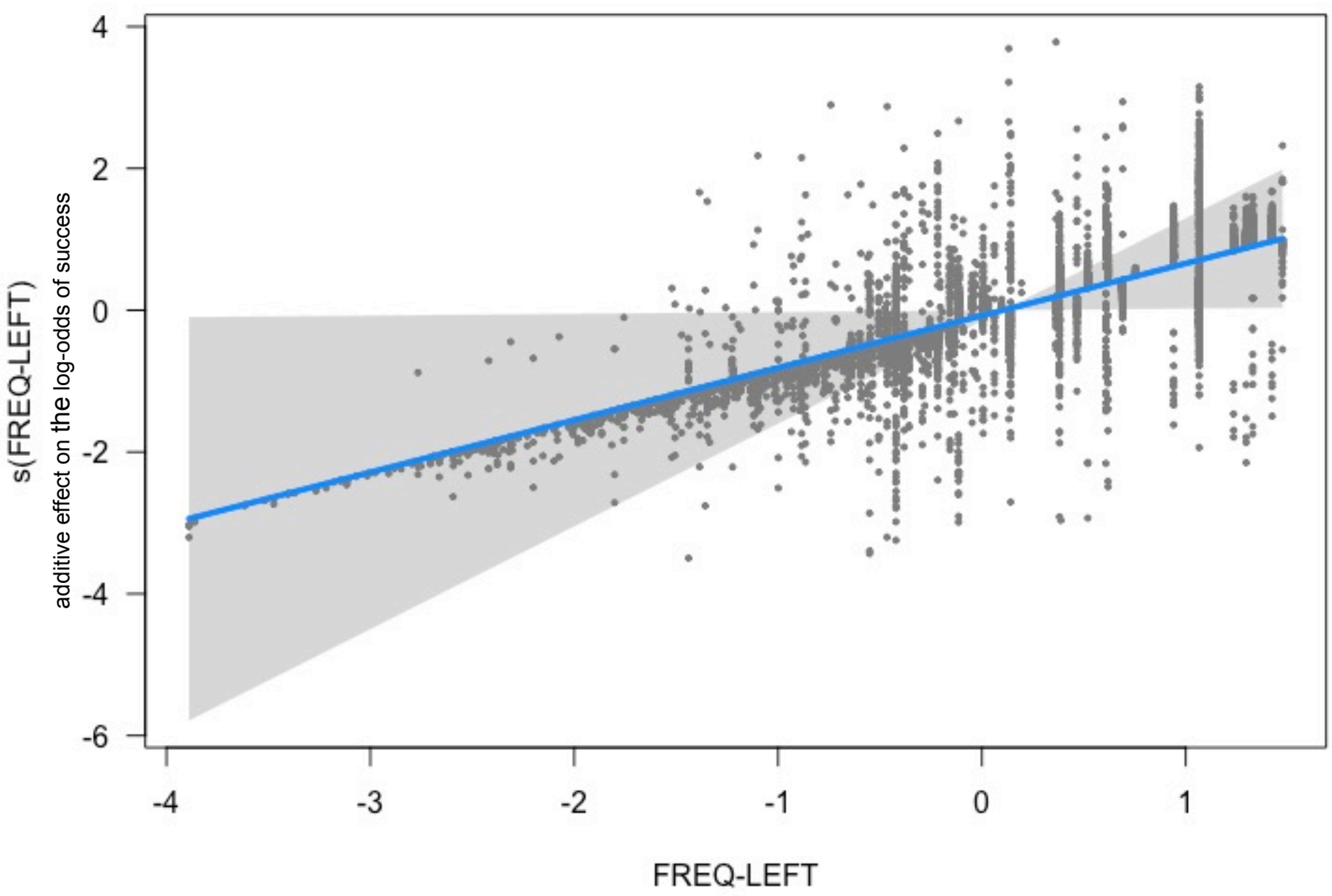
$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

6.1 Le modèle final - Les fonctions de lissage

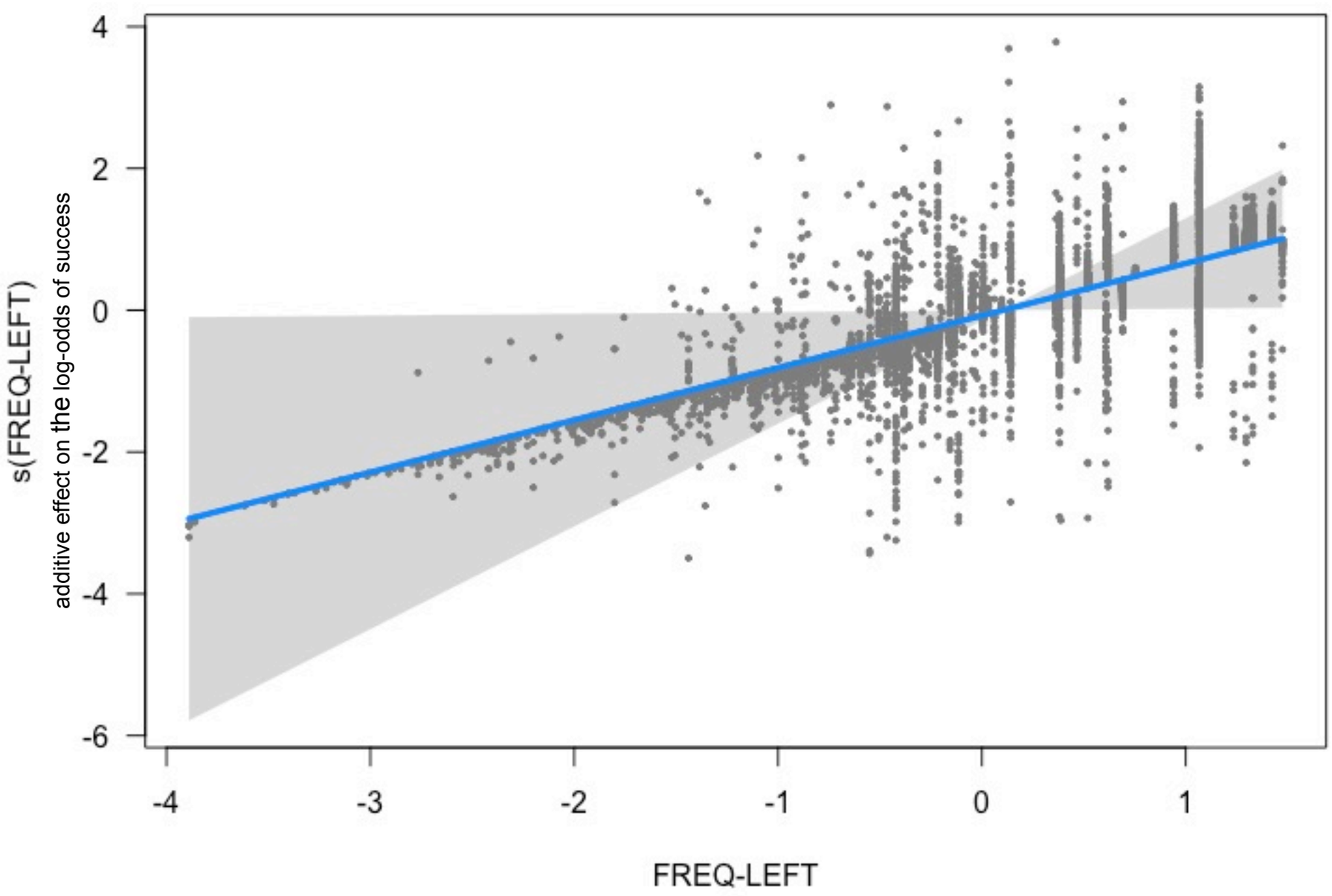
$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

L'augmentation de FREQ-G détermine une plus forte probabilité d'avoir une liaison réalisée



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

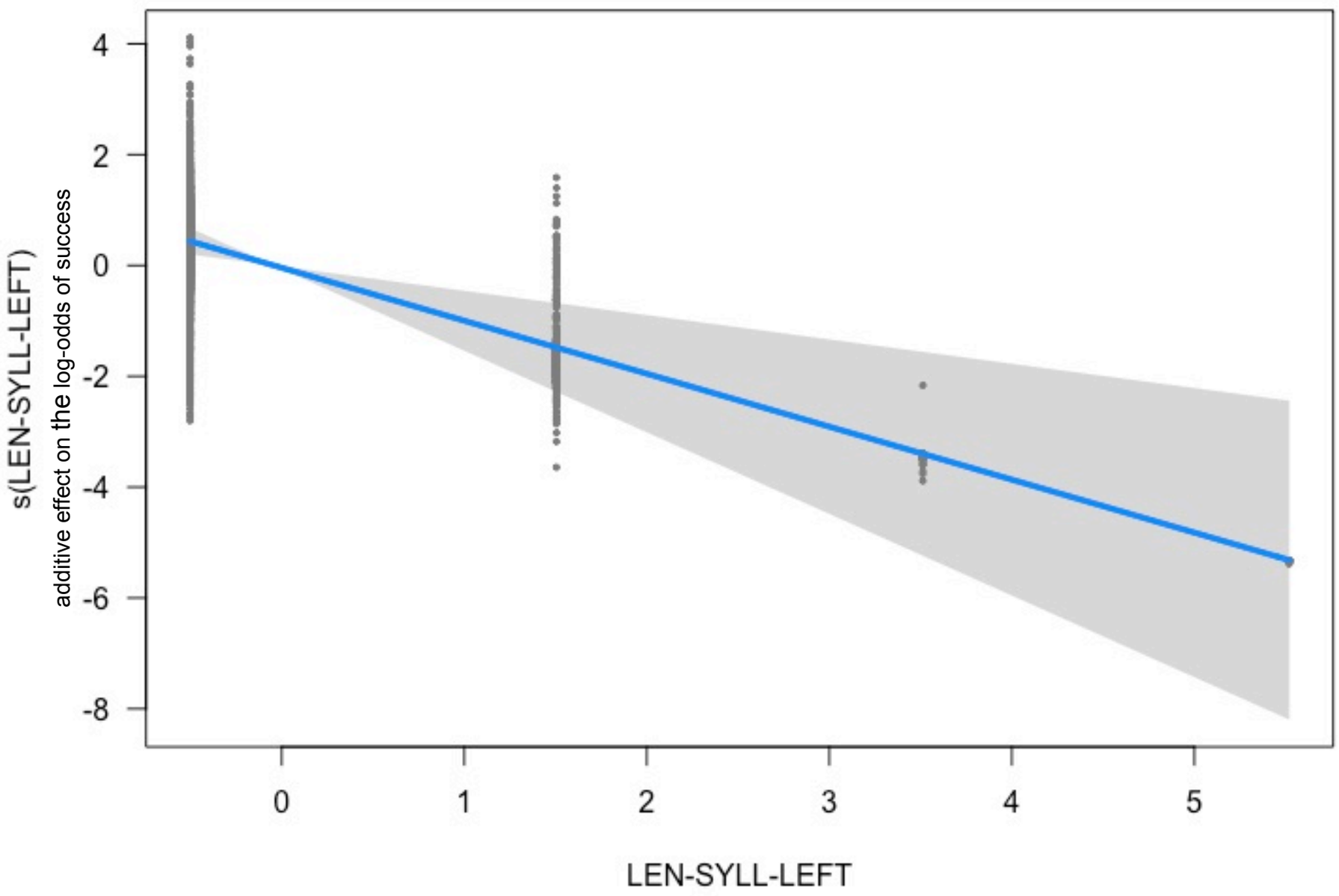
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

6.1 Le modèle final - Les fonctions de lissage

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

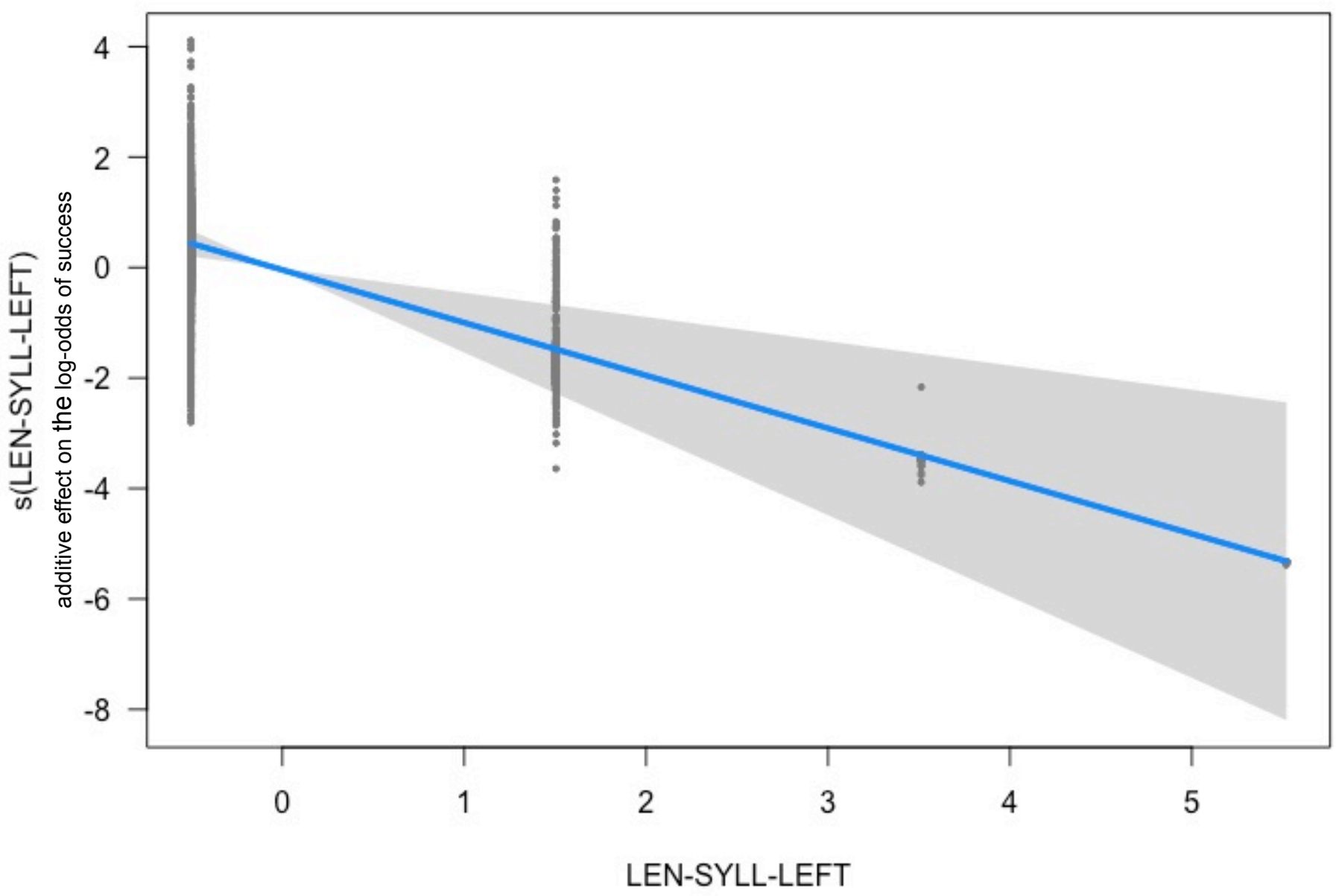
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

L'augmentation de SYLL-G détermine une probabilité plus faible d'avoir une liaison réalisée



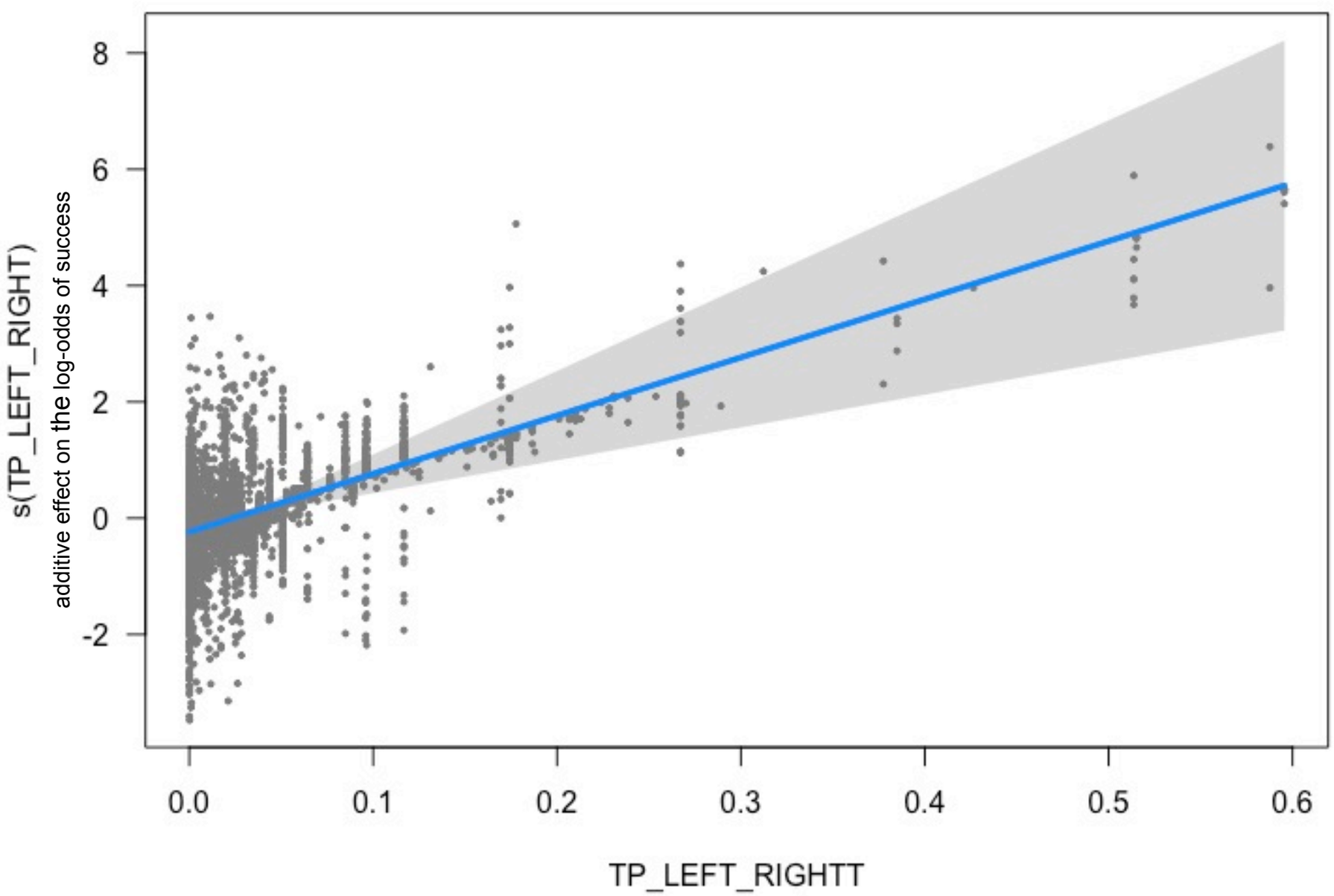
$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

6.1 Le modèle final - Les fonctions de lissage

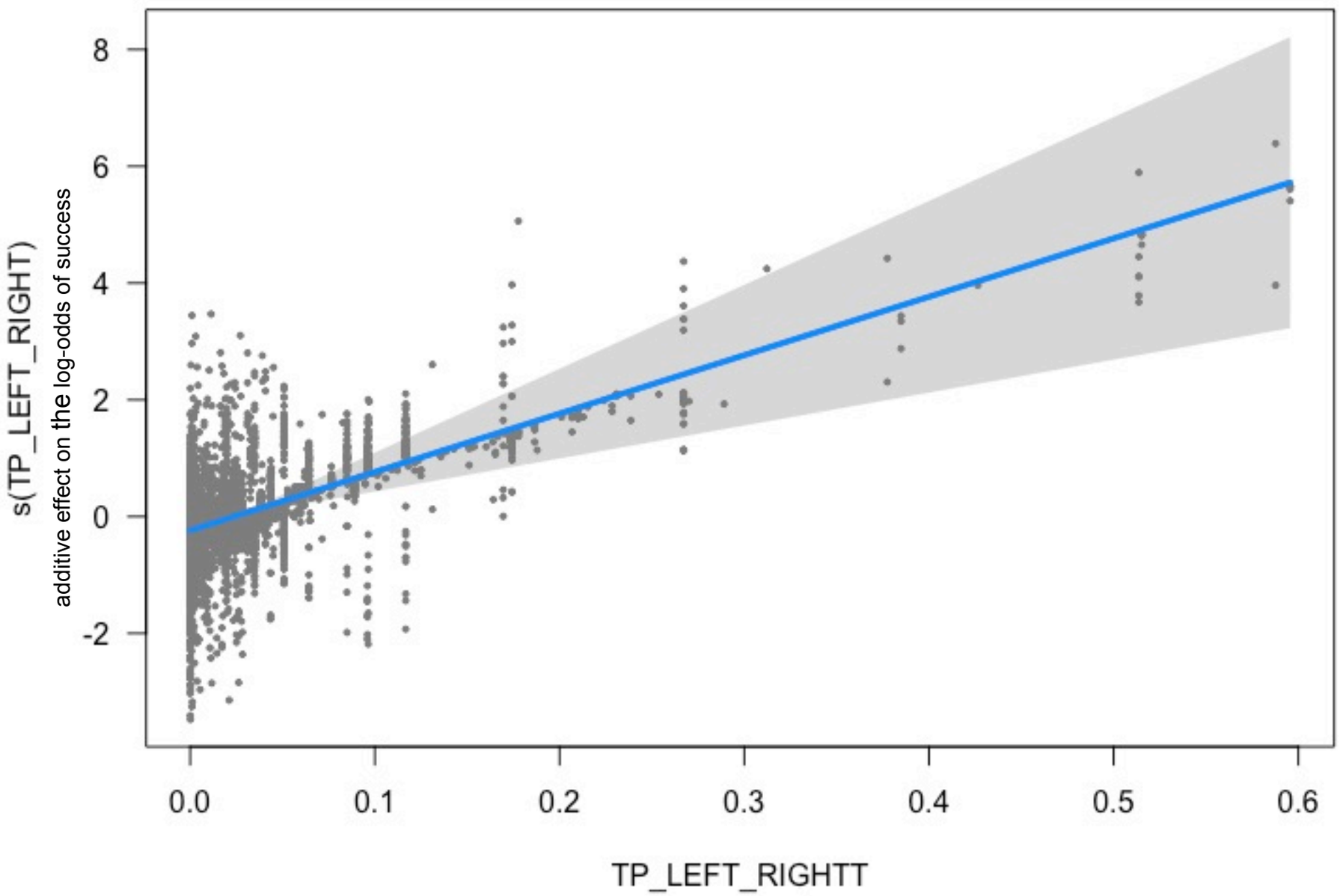
$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

L'augmentation de TP.BIGRAM détermine une plus forte probabilité d'avoir une liaison réalisée



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

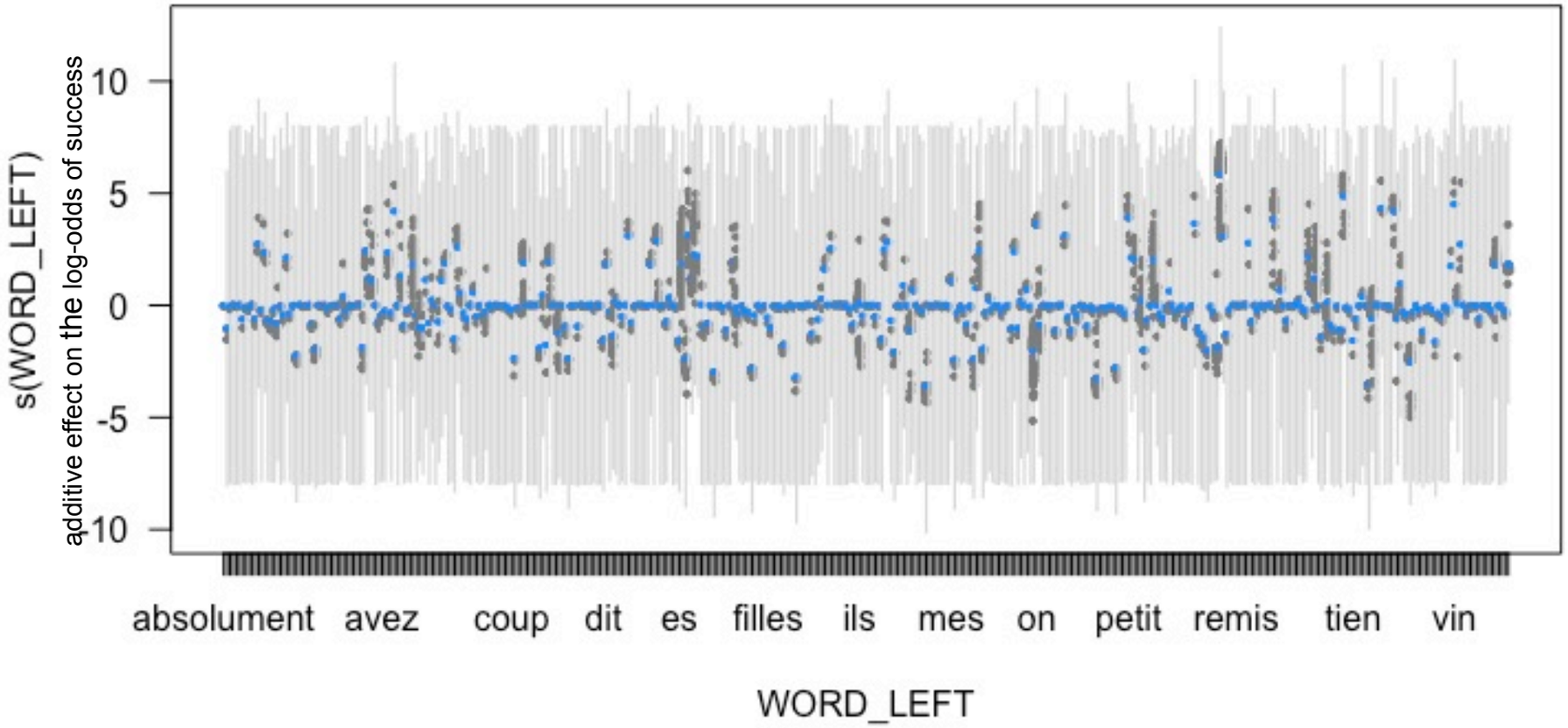
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

6.1 Le modèle final - Les fonctions de lissage

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

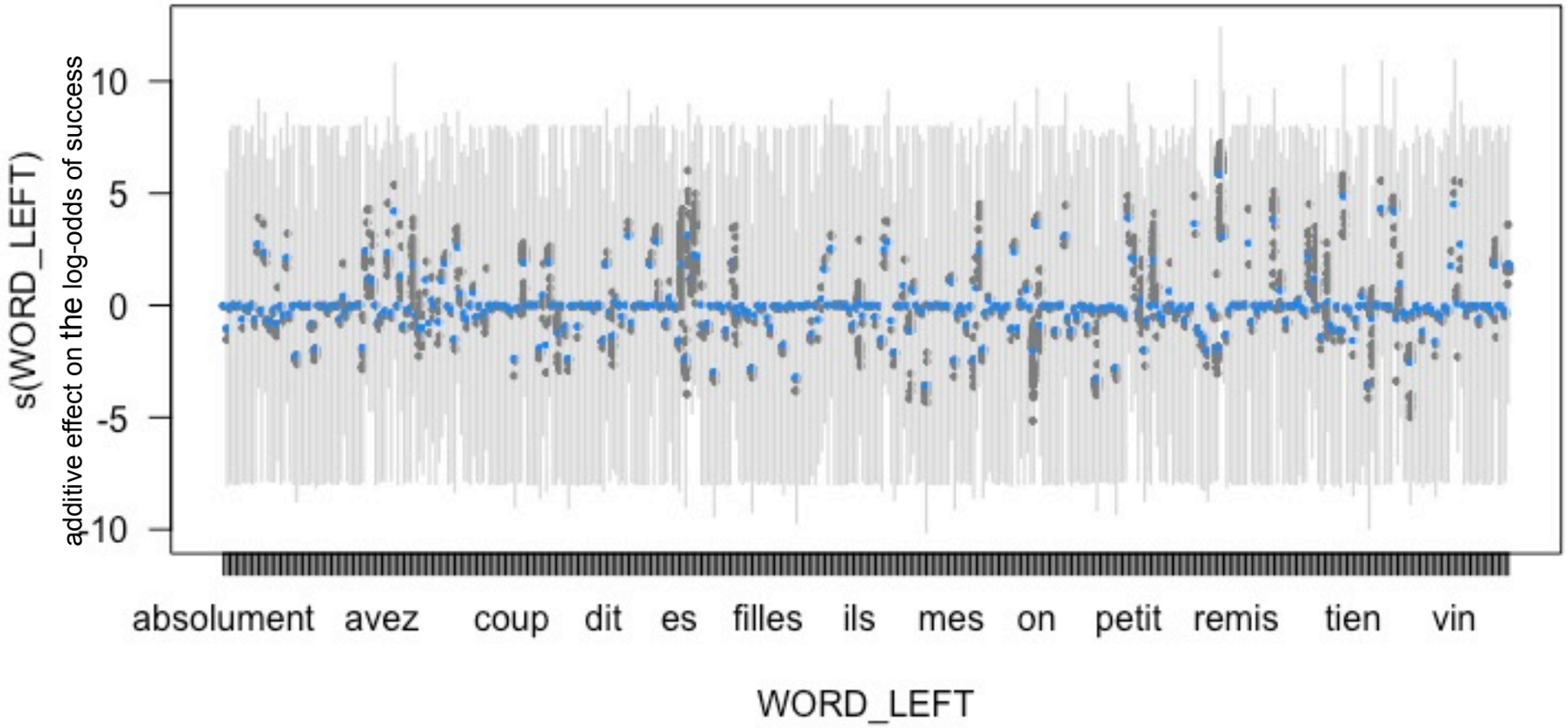
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

La probabilité de réalisation d'une liaison est spécifique à chaque valeur de MOT-G



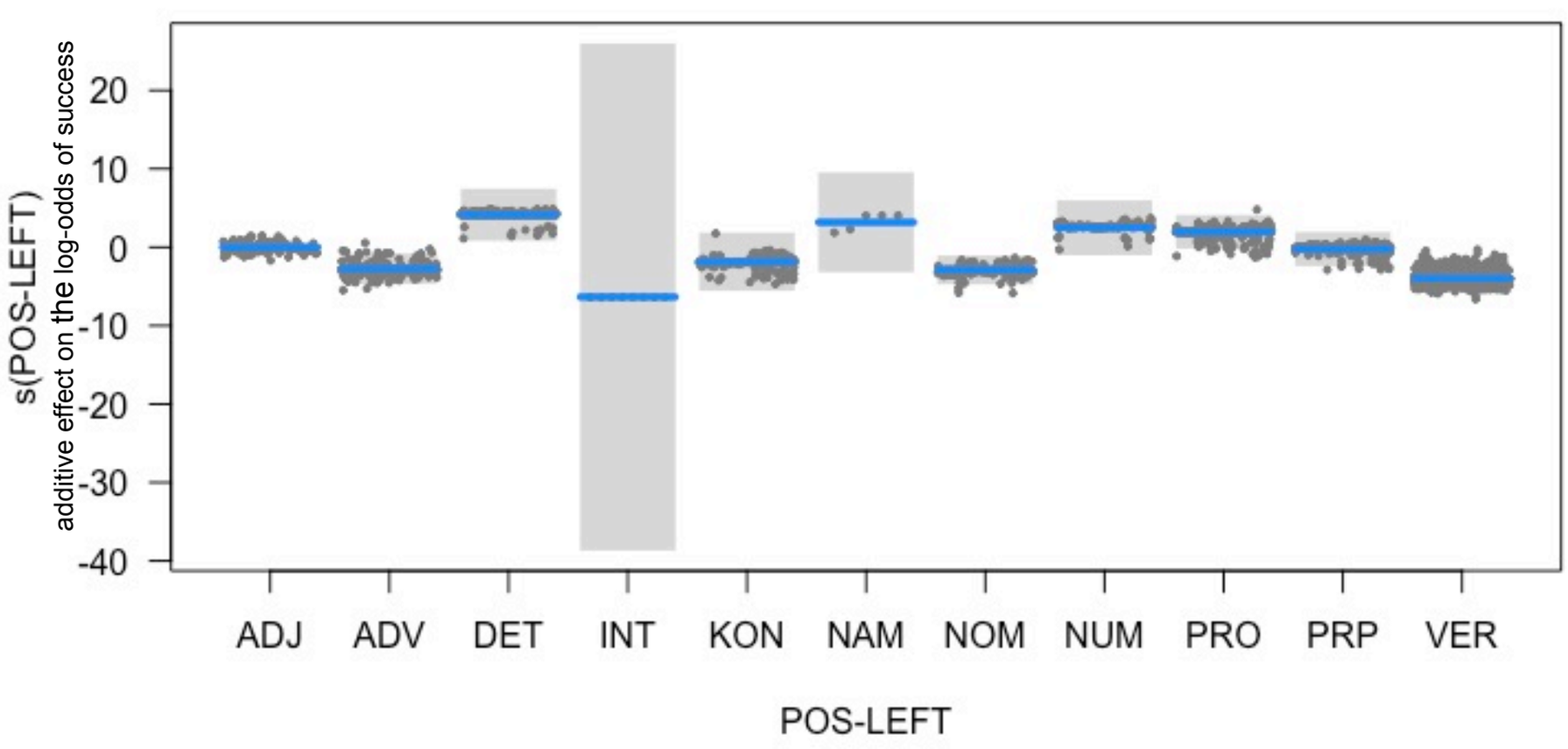
$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



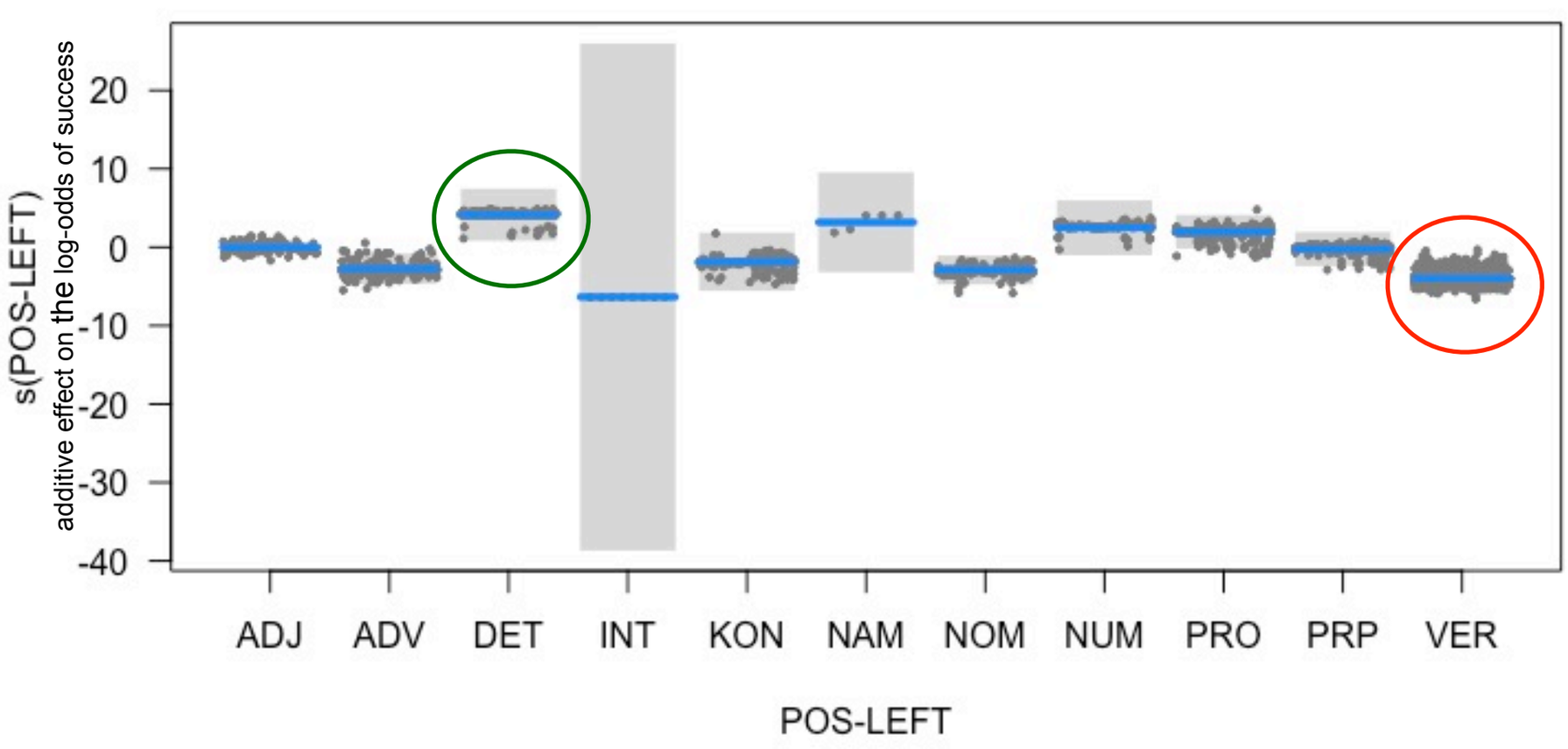
ADJ=adjective
ADV=adverb
DET=determinant
INT=interjection
KON=conjunction
NAM=proper name
NOM=noun
NUM=numeral
PRO=pronoun
PRP=preposition
VER=verb

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

La probabilité de réalisation d'une liaison est liée à chaque valeur de POS-G

POS-G qui favorise mieux la réalisation de la liaison : DET

POS-G qui favorise moins la réalisation de la liaison : VER



ADJ=adjective
ADV=adverb
DET=determinant
INT=interjection
KON=conjunction
NAM=proper name
NOM=noun
NUM=numeral
PRO=pronoun
PRP=preposition
VER=verb

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

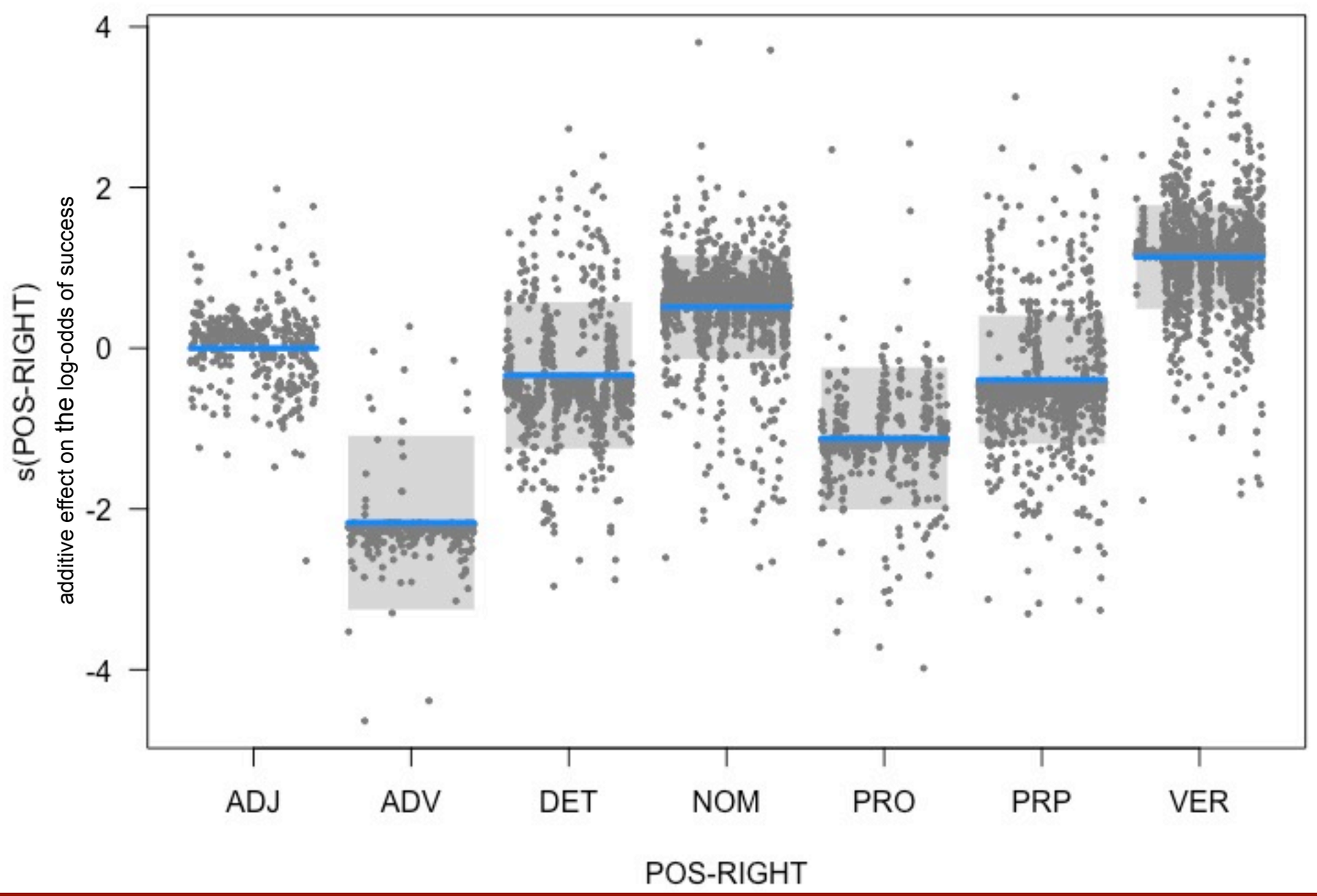
6.1 Le modèle final - Les fonctions de lissage

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



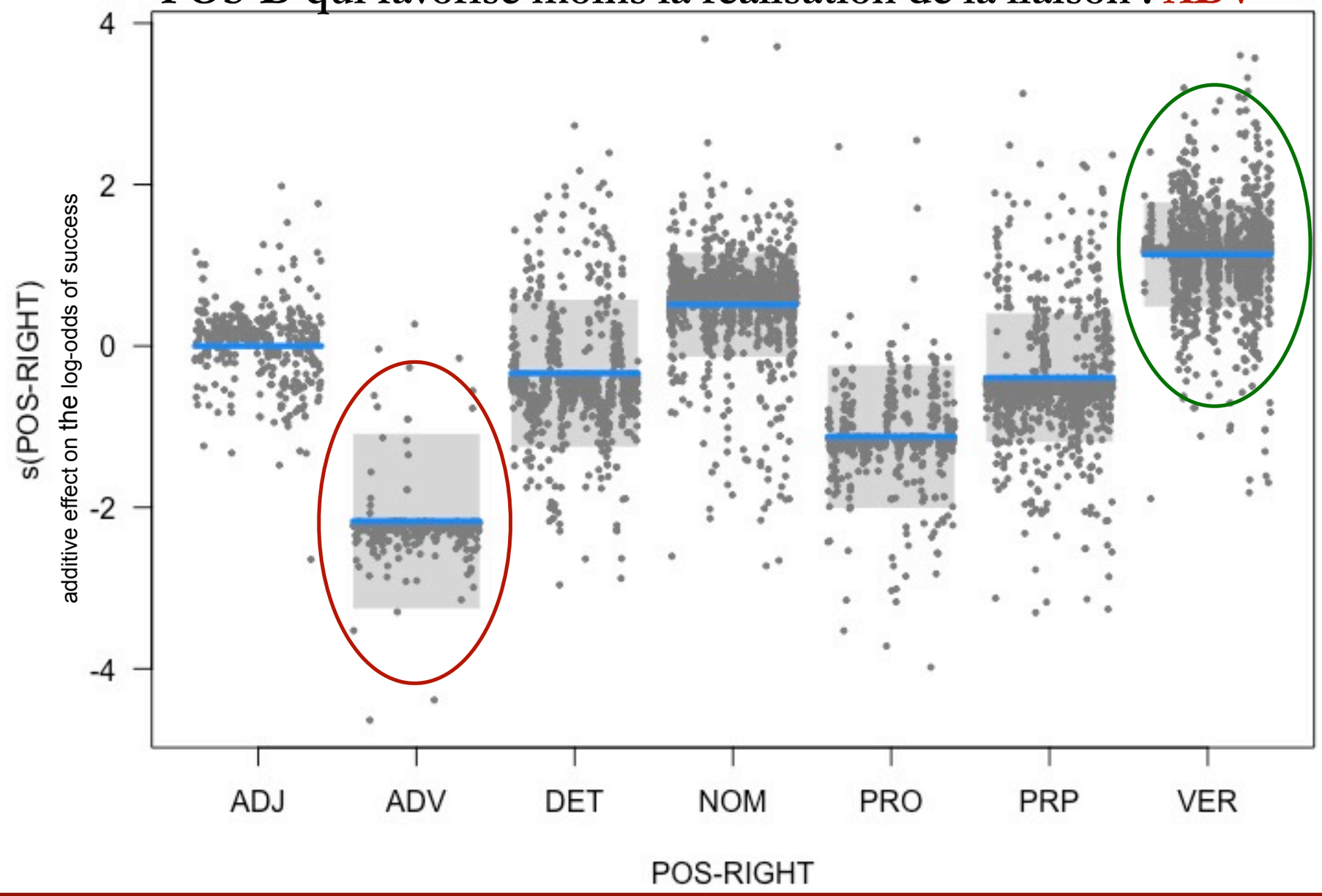
ADJ=adjective
ADV=adverb
DET=determinant
INT=interjection
KON=conjunction
NAM=proper name
NOM=noun
NUM=numeral
PRO=pronoun
PRP=preposition
VER=verb

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

La probabilité de réalisation d'une liaison est liée à chaque valeur de POS-D

POS-D qui favorise mieux la réalisation de la liaison : VER

POS-D qui favorise moins la réalisation de la liaison : ADV



ADJ=adjective
ADV=adverb
DET=determinant
INT=interjection
KON=conjunction
NAM=proper name
NOM=noun
NUM=numeral
PRO=pronoun
PRP=preposition
VER=verb

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

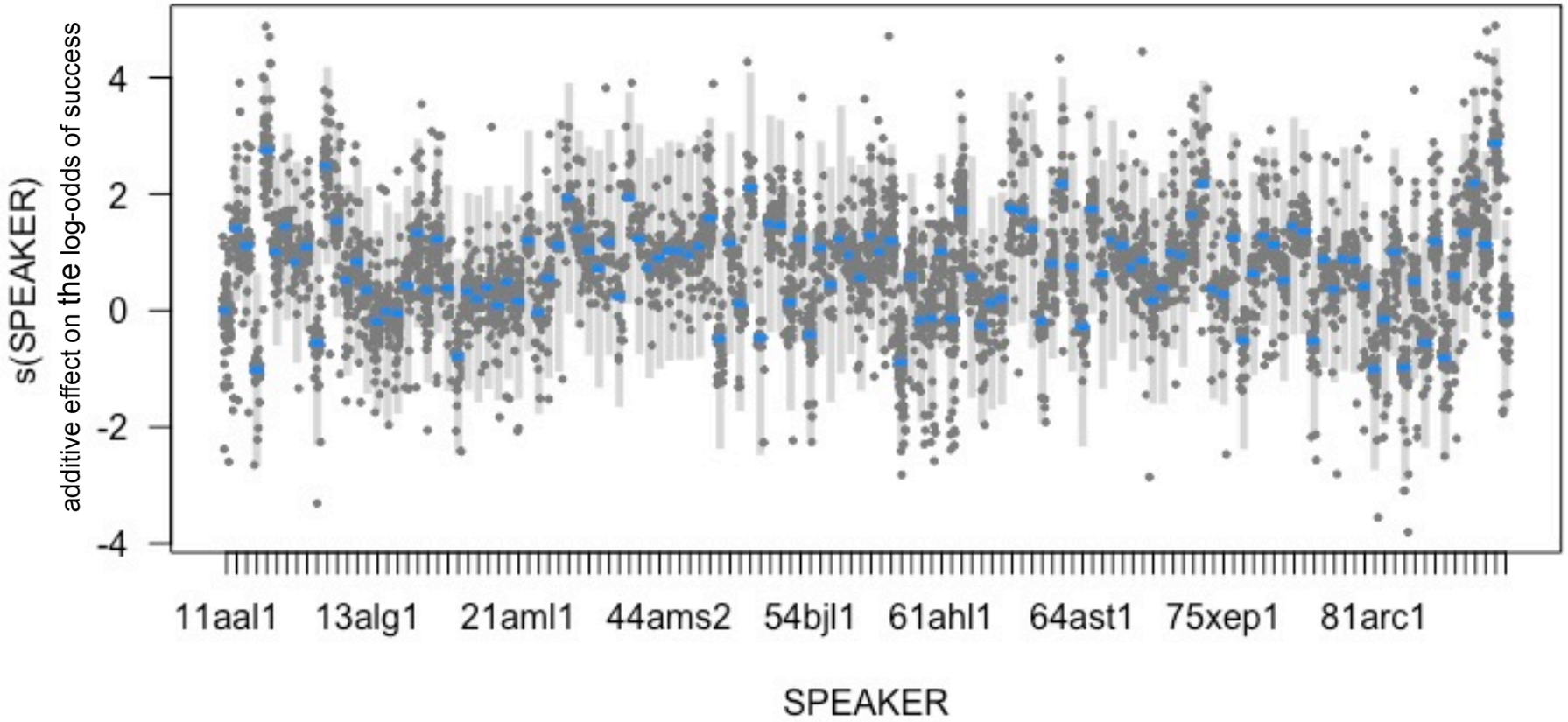
6.1 Le modèle final - Les fonctions de lissage

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

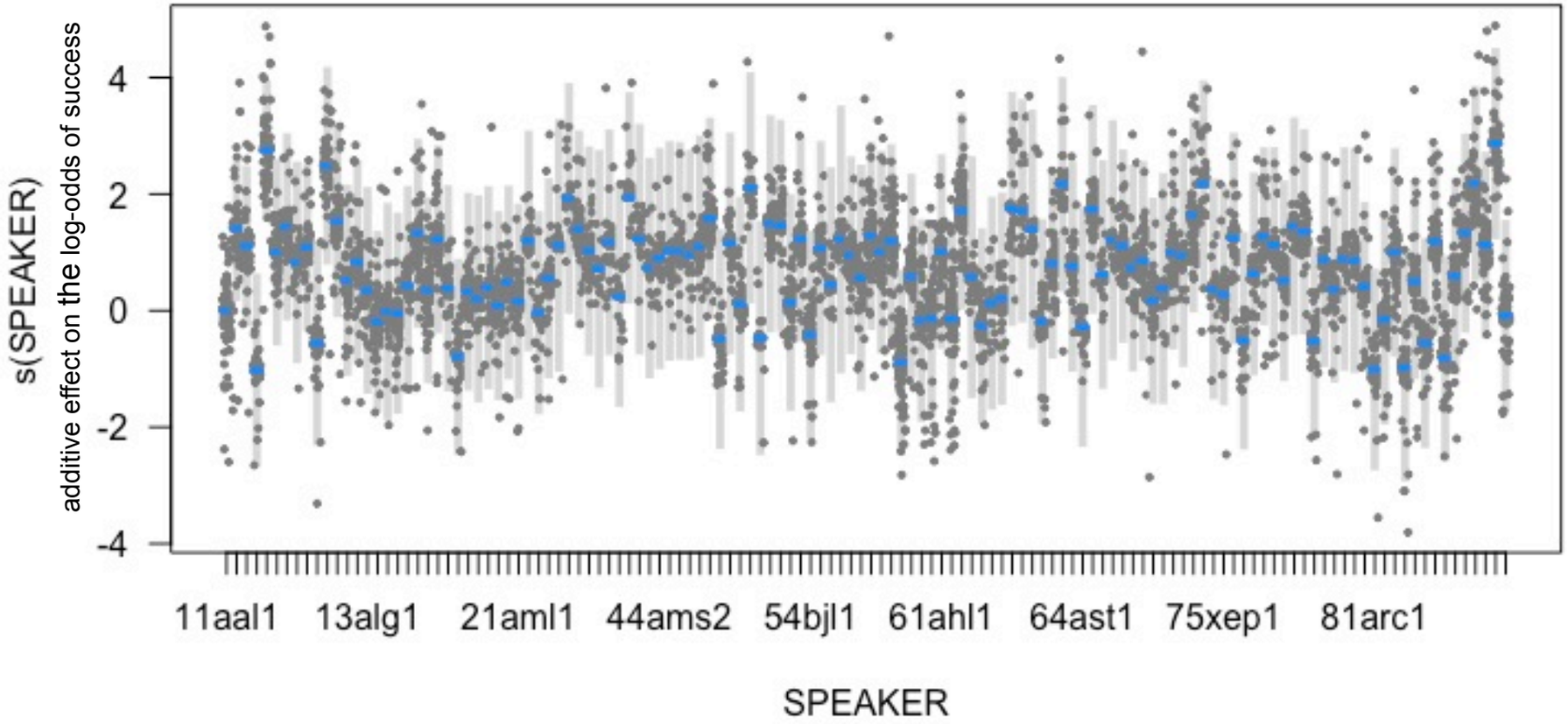
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ{-}G) + s(SYLL{-}G) + s(TP.BIGRAM) + s(MOT{-}G) +$$

$$+ s(POS{-}G) + s(POS{-}D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

La probabilité de réalisation d'une liaison est spécifique à chaque valeur de LOCUTEUR



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

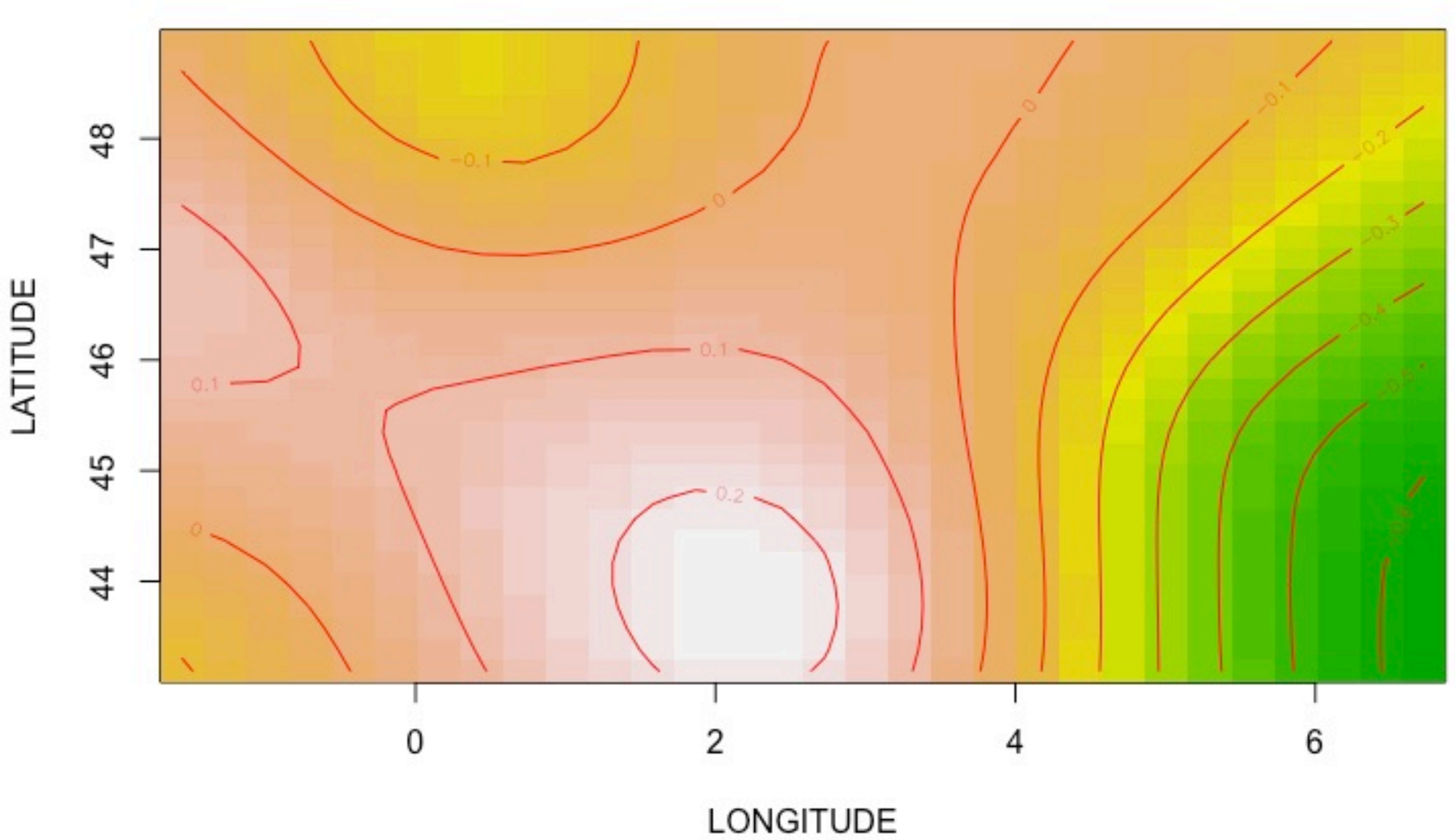
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

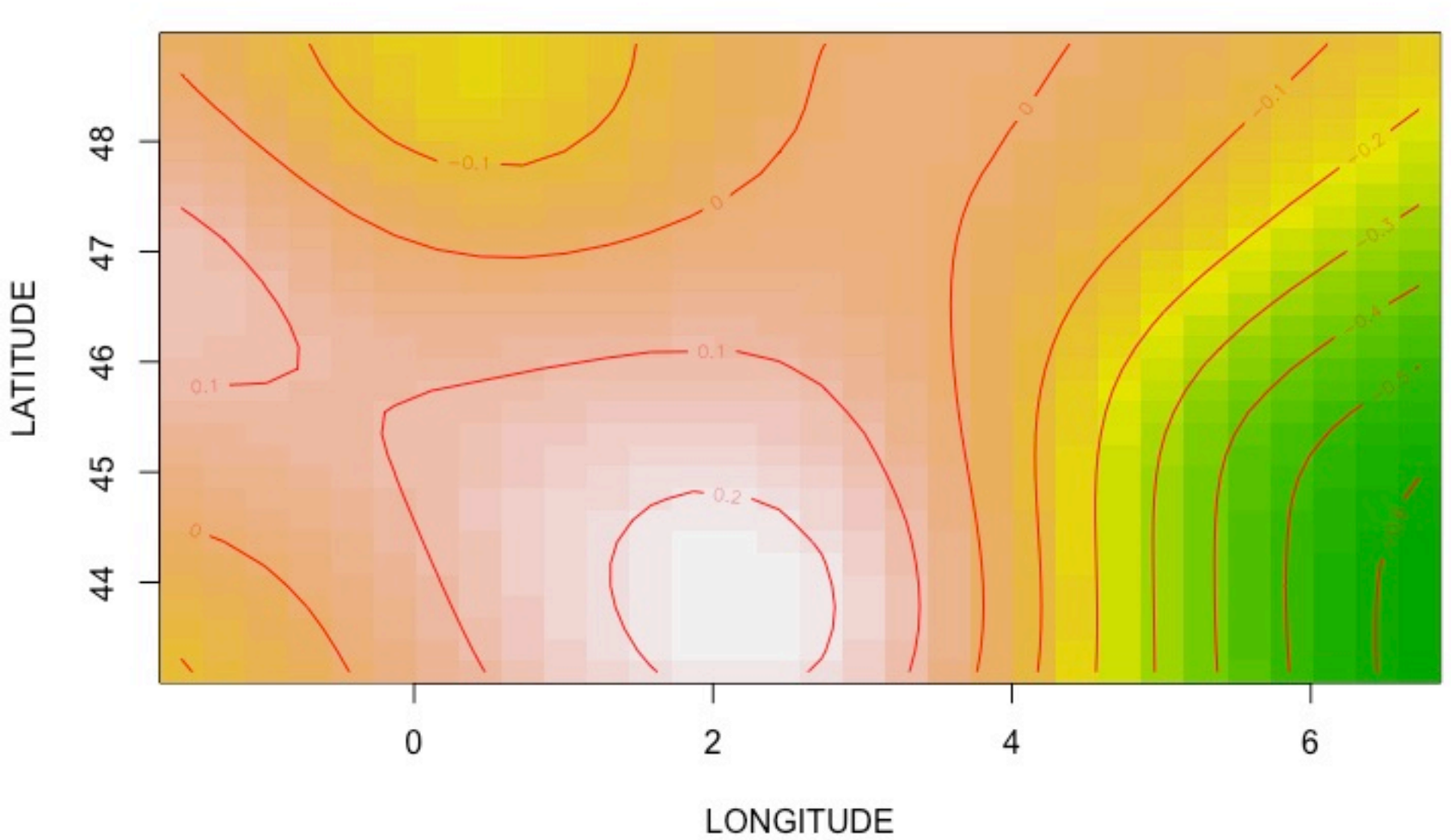
$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

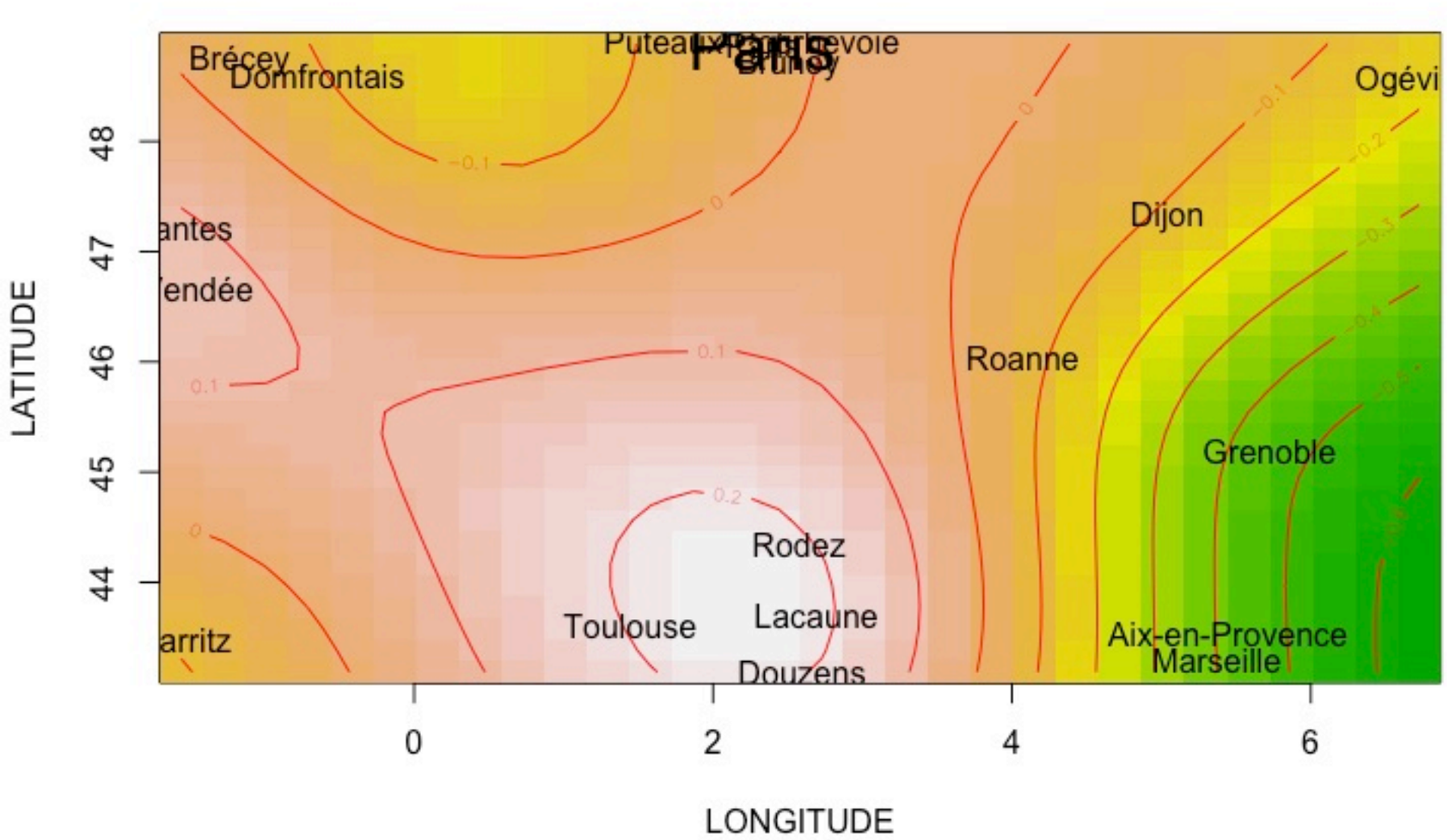
La probabilité de réalisation d'une liaison par rapport à la LONGITUDE et LATITUDE du locuteur



$$model_{\text{FINAL}} = -1.915713 + s(AGE) + s(FREQ-G) + s(SYLL-G) + s(TP.BIGRAM) + s(MOT-G) +$$

$$+ s(POS-G) + s(POS-D) + s(LOCUTEUR) + s(LONGITUDE, LATITUDE)$$

La probabilité de réalisation d'une liaison par rapport à la LONGITUDE et LATITUDE du locuteur



- Objectif du travail
- La liaison - une petite présentation
- Un corpus phonologique: PFC
- Les outils statistiques (RL et GAM)
- Le model final et le résultats
- Limites du travail et perspectives futures

7. Conclusions

7. Conclusions

- Nous avons crée un modèle prédictif qui explique plus que 83% de la variabilité de la liaison en français

7. Conclusions

- Nous avons crée un modèle prédictif qui explique plus que 83% de la variabilité de la liaison en français

- Parmi les variables explicatives, le mot à gauche (et son POS) semble être la variable la plus explicative du phénomène liaison, suivi par des variables distributionnelles comme la fréquence du mot à gauche et des variables socio-linguistiques comme l'âge du locuteur

- Nous avons crée un modèle prédictif qui explique plus que 83% de la variabilité de la liaison en français

- Parmi les variables explicatives, le mot à gauche (et son POS) semble être la variable la plus explicative du phénomène liaison, suivi par des variables distributionnelles comme la fréquence du mot à gauche et des variables socio-linguistiques comme l'âge du locuteur

- Le modèle est prédictif. Il est possible prédire la probabilité de liaison de données inconnues : P[(*Liaison=oui*) | M. Dupont de Bordeaux de 70 ans face à la construction <pas encore>]

- Nous avons crée un modèle prédictif qui explique plus que 83% de la variabilité de la liaison en français

- Parmi les variables explicatives, le mot à gauche (et son POS) semble être la variable la plus explicative du phénomène liaison, suivi par des variables distributionnelles comme la fréquence du mot à gauche et des variables socio-linguistiques comme l'âge du locuteur

- Le modèle est prédictif. Il est possible prédire la probabilité de liaison de données inconnues : P[(*Liaison=oui*) | M. Dupont de Bordeaux de 70 ans face à la construction <pas encore>]

- Le modèle incorpore aussi des **effets aléatoires** (pas illustrés aujourd'hui) pour les variables liés aux **locuteurs** et aux **mots** (Baayen, 2014).

- Nous avons crée un modèle prédictif qui explique plus que 83% de la variabilité de la liaison en français

- Parmi les variables explicatives, le mot à gauche (et son POS) semble être la variable la plus explicative du phénomène liaison, suivi par des variables distributionnelles comme la fréquence du mot à gauche et des variables socio-linguistiques comme l'âge du locuteur

- Le modèle est prédictif. Il est possible prédire la probabilité de liaison de données inconnues : P[(*Liaison=oui*) | M. Dupont de Bordeaux de 70 ans face à la construction <pas encore>]

- Le modèle incorpore aussi des **effets aléatoires** (pas illustrés aujourd'hui) pour les variables liés aux **locuteurs** et aux **mots** (Baayen, 2014).

  - Nous n'avons pas enregistré le comportement de liaison chez tous les locuteurs francophones, mais seulement chez un nombre restreint (192 locuteurs), pris aléatoirement.

- Nous avons crée un modèle prédictif qui explique plus que 83% de la variabilité de la liaison en français

- Parmi les variables explicatives, le mot à gauche (et son POS) semble être la variable la plus explicative du phénomène liaison, suivi par des variables distributionnelles comme la fréquence du mot à gauche et des variables socio-linguistiques comme l'âge du locuteur

- Le modèle est prédictif. Il est possible prédire la probabilité de liaison de données inconnues : P[(*Liaison=oui*) | M. Dupont de Bordeaux de 70 ans face à la construction <pas encore>]

- Le modèle incorpore aussi des **effets aléatoires** (pas illustrés aujourd'hui) pour les variables liés aux **locuteurs** et aux **mots** (Baayen, 2014).

  - Nous n'avons pas enregistré le comportement de liaison chez tous les locuteurs francophones, mais seulement chez un nombre restreint (192 locuteurs), pris aléatoirement.
  - Il y aura sûrement des différences de comportement de liaison entre un locuteur et l'autre et même chez le même locuteur. Cette variation est due à des facteurs incontrôlables (sujets plus sensibles à la langue, tension/émotion lorsque de la registration, etc).

- Nous avons crée un modèle prédictif qui explique plus que 83% de la variabilité de la liaison en français

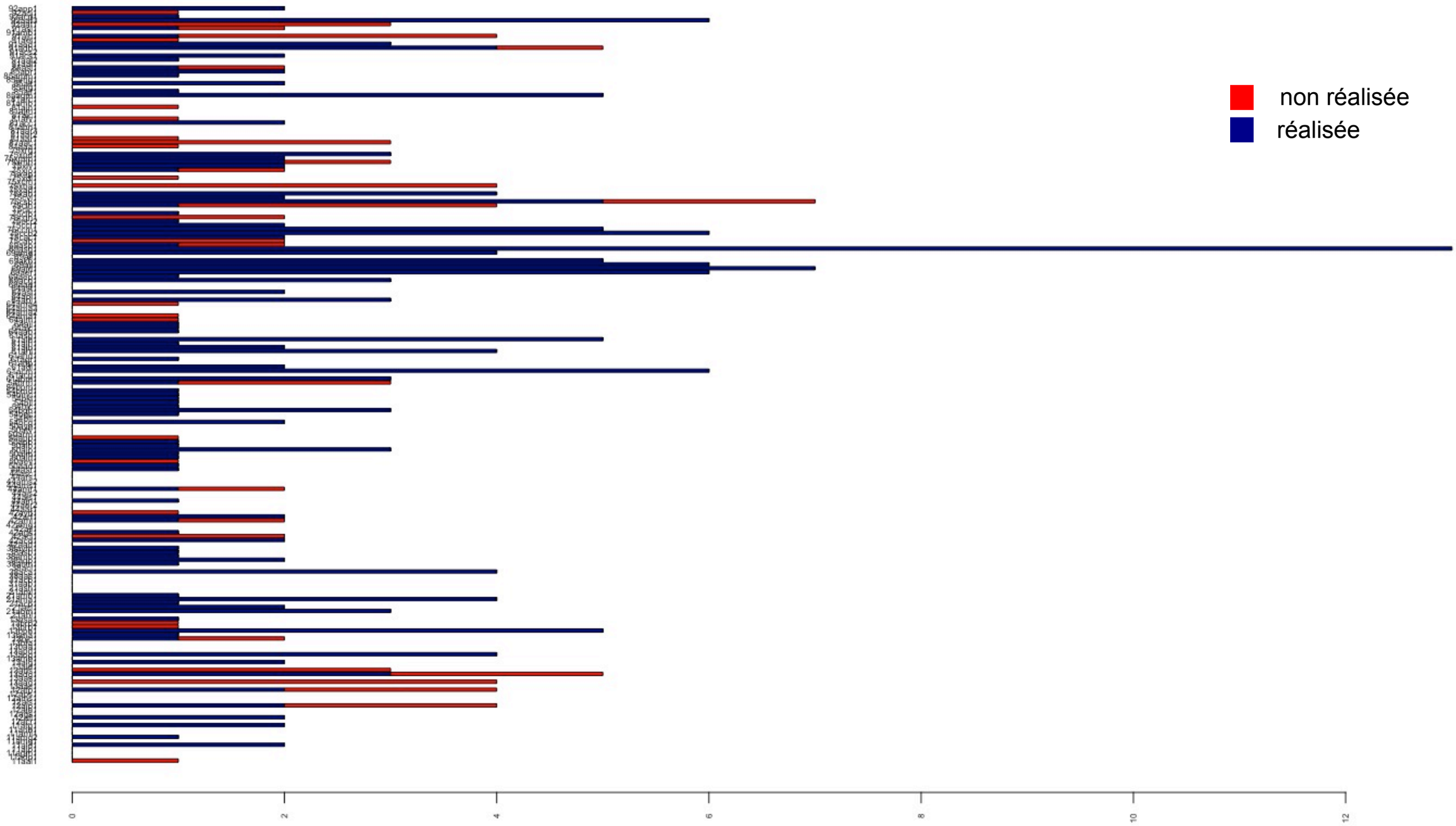
- Parmi les variables explicatives, le mot à gauche (et son POS) semble être la variable la plus explicative du phénomène liaison, suivi par des variables distributionnelles comme la fréquence du mot à gauche et des variables socio-linguistiques comme l'âge du locuteur

- Le modèle est prédictif. Il est possible prédire la probabilité de liaison de données inconnues : P[(*Liaison=oui*) | M. Dupont de Bordeaux de 70 ans face à la construction <pas encore>]

- Le modèle incorpore aussi des **effets aléatoires** (pas illustrés aujourd'hui) pour les variables liés aux **locuteurs** et aux **mots** (Baayen, 2014).

  - Nous n'avons pas enregistré le comportement de liaison chez tous les locuteurs francophones, mais seulement chez un nombre restreint (192 locuteurs), pris aléatoirement.
  - Il y aura sûrement des différences de comportement de liaison entre un locuteur et l'autre et même chez le même locuteur. Cette variation est due à des facteurs incontrôlables (sujets plus sensibles à la langue, tension/émotion lorsque de la registration, etc).
  - Grâce à l'introduction d'un effet aléatoire spécifique à chaque sujet (et à chaque mot) le modèle a pris en comptes ces différences.

Distribution de <est un> chez les locuteurs



7. Limites du travail et perspectives futures

- Une limite de ce modèle est le fait de ne pas prendre en compte la dimension syntaxique (bien que l'on considère les informations de POS des mots à gauche et à droite, on ne possède pas d'information sur les syntagmes dans lesquels ces mots sont impliqués)

- Une limite de ce modèle est le fait de ne pas prendre en compte la dimension syntaxique (bien que l'on considère les informations de POS des mots à gauche et à droite, on ne possède pas d'information sur les syntagmes dans lesquels ces mots sont impliqués)

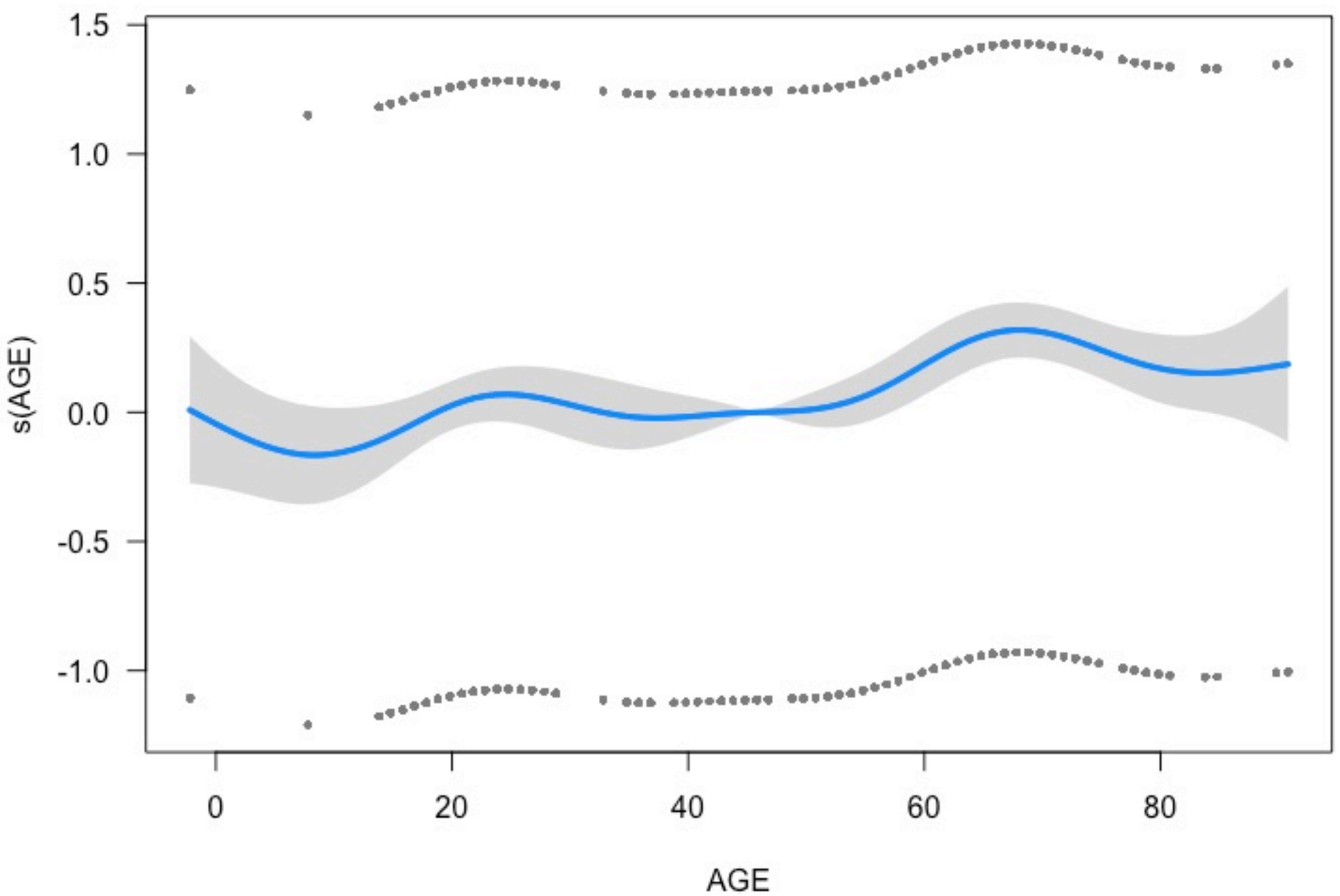
- Une limite ultérieure concerne les aspects phonotactiques : les informations sur les phonèmes à gauche et à droite pourraient en effet enrichir le modèle

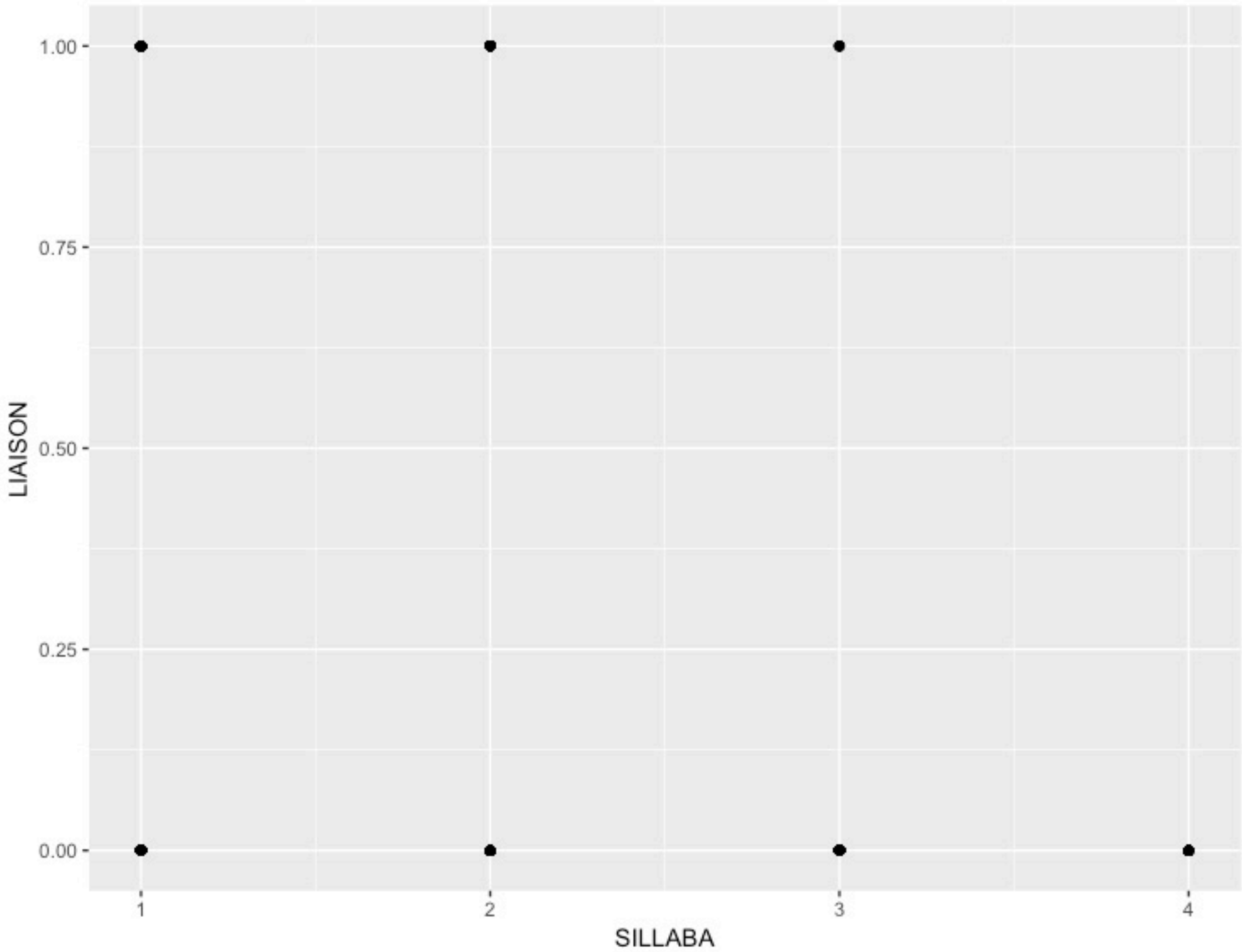
- Une limite de ce modèle est le fait de ne pas prendre en compte la dimension syntaxique (bien que l'on considère les informations de POS des mots à gauche et à droite, on ne possède pas d'information sur les syntagmes dans lesquels ces mots sont impliqués)

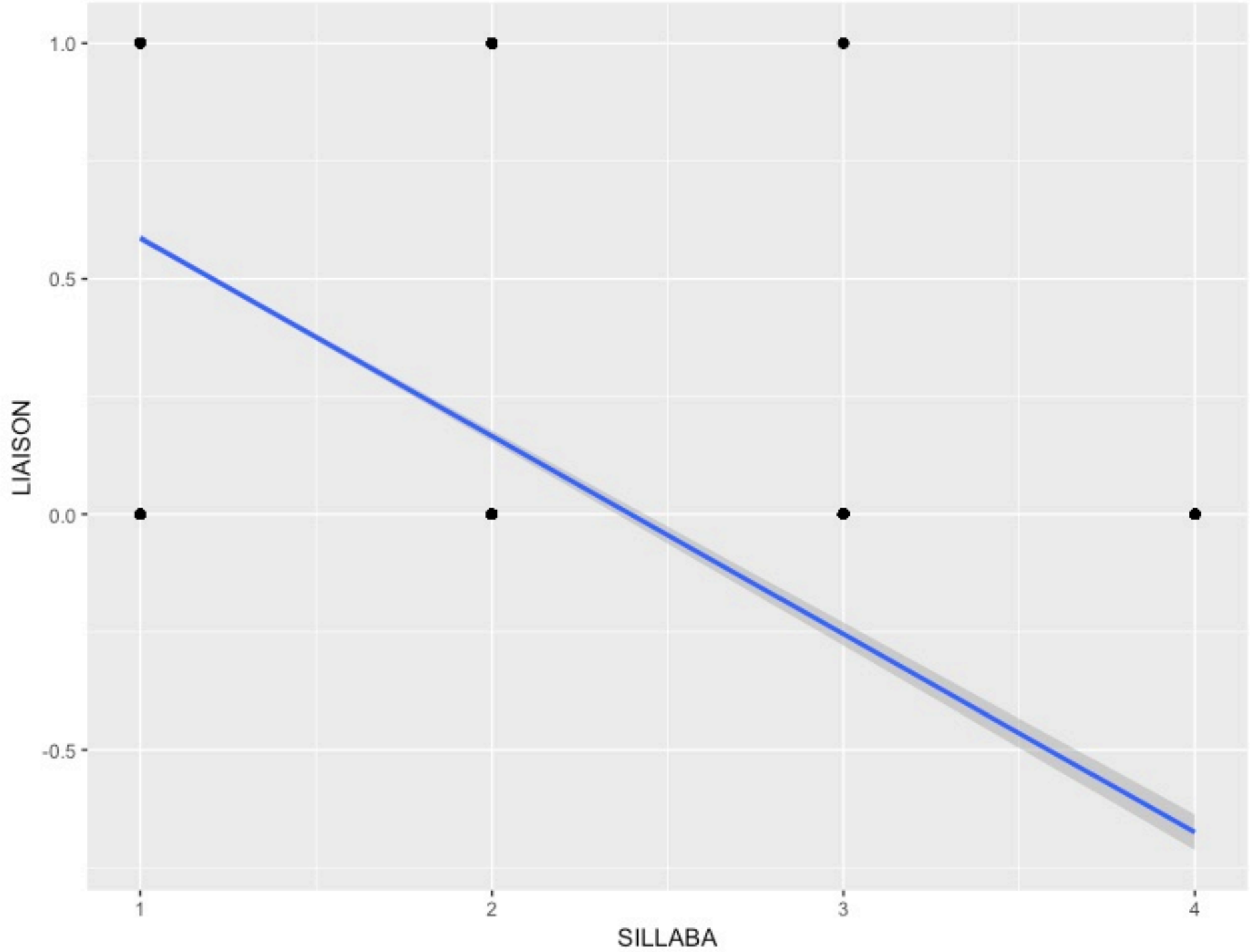
- Une limite ultérieure concerne les aspects phonotactiques : les informations sur les phonèmes à gauche et à droite pourraient en effet enrichir le modèle

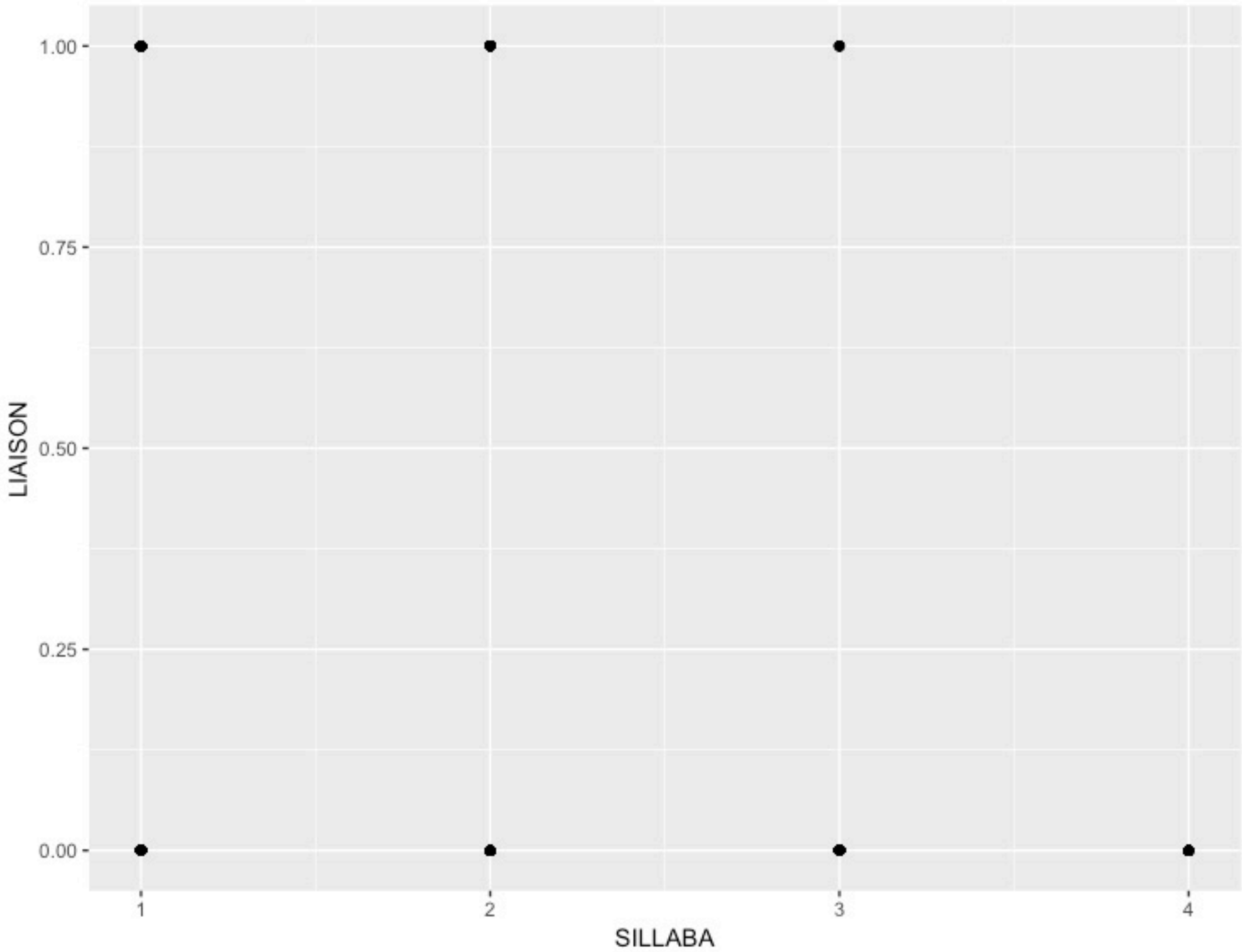
- Considérer aussi la dimension holophrastique d'occurrence de la liaison

MERCI


1.5
1.0
0.5
0.0
-0.5
-1.0
s(AGE)
0
20
40
60
80
AGE


1.00
0.75
0.50
0.25
0.00
LIAISON
1
2
3
4
SILLABA


1.0
0.5
0.0
-0.5
LIAISON
1
2
3
4
SILLABA


1.00
0.75
0.50
0.25
0.00
LIAISON
1
2
3
4
SILLABA

1. La liaison en français