Projet SAS

ANALYSE DES FACTEURS ASSOCIEES AU POIDS DES NOURRISSONS

Etude à partir des données collectées au centre médical de Baystate dans le Massachusetts pendant l'année 1986

Rapport réalisé par :

Rakotondrazaka Diamondra

Royer Mathis

Résumé :

Le faible poids de naissance est un événement qui intéresse les médecins depuis plusieurs années en raison du taux de mortalité et d'anomalies infantiles très élevés chez les nourrissons de faible poids.

Il s'agit d'une enquête concernant les facteurs de risques associés au faible poids de naissance de nourrissons réalisée au centre médical de Baystate dans le Massachusetts pendant l'année 1986 avec 189 femmes interrogées. Le comportement d'une femme pendant la grossesse (régime alimentaire, habitudes tabagiques ...) peut altérer de façon importante les chances de mener correctement la grossesse à terme et par conséquent de donner naissance à un enfant de poids normal. Le fichier de données contient les informations sur 189 femmes venant consulter dans le centre médical. On considère qu'un enfant a un faible poids de naissance si celui-ci est inférieur à 2500 g.

Voici la liste des variables de notre étude :

Description	**Unité** ou **Codage**	**Variable**
Age de la mère	Années	AGE
Poids de la mère lors du dernier cycle menstruel	Livres	LWT
Race de la mère	1=Blanche ; 2=Noire ; 3=Autre	RACE
Tabagisme durant la grossesse	Oui=1 ; Non=0	SMOKE
Nombre d'antécédents de prématurité	0=Non ; 1=Un ; 2=Deux ; etc ...	PTL
Antécédents d'hypertension	Oui=1 ; Non=0	HT
Présence d'irritabilité utérine	Oui=1 ; Non=0	UI
Nombre de visites à un médecin durant le premier trimestre de la grossesse	0=Aucune ; 1=Une ; etc...	FVT
Poids de naissance	Grammes	BWT
Poids de naissance inférieur ou égal à 2500 g	Oui=1 ; Non=0	LOW

Annexe 1 : liste des variables de l'étude.

Table des matières

Résumé : 2
Liste des Figures : 4
Liste des Tableaux : 4
I Traitement des données : 5
II Analyse univariée : 6
III Analyse bivariée : 10
1) Variables qualitatives : 10
a) Tests d'indépendance de Fisher. 10
b) Calcul des odds ratio. 11
c) Représentations graphiques de la distribution du poids de naissance des bébés en fonction des facteurs à modalités a priori significatifs. 12
2) Variables quantitatives : 14
a) Tests de normalité à 2 échantillons d'une variable. 14
b) Tests non-paramétriques de comparaison. 14
IV Analyse multivariée : 18
1) Régression logistique multiple. 18
2) Analyse en correspondance multiples : 19
Conclusion : 22
Références : 23
Annexes : 23
Codes sous SAS : 24
Codes sous R : 30

Liste des Figures :

Figure 1 : boxplots de FVT et BWT......6
Figure 2 : boxplots de LWT, PTL et AGE......7
Figure 3 : diagrammes en bâtons des variables qualitatives......9
Figure 4 : distribution de BWT en fonction de HT......12
Figure 5 : distribution de BWT en fonction de UI......13
Figure 6 : distribution de BWT en fonction de SMOKE......13
Figure 7 : nuage de points de BWT en fonction de LWT......16
Figure 8 : nuage de points de BWT en fonction de PTL......17
Figure 9 : probabilité d'avoir LOW = "oui" en fonction de LWT......18
Figure 10 : probabilité d'avoir LOW = "oui" en fonction de PTL......19
Figure 11 : représentation graphique de l'ACM......22

Liste des Tableaux :

Tableau 1 : statistiques descriptives des variables quantitatives......6
Tableau 2 : statistiques descriptives des variables qualitatives......8
Tableau 3 : Résultats du test exact de Fisher......10
Tableau 4 : Résultats du test de Fisher......10
Tableau 5 : rapport des cotes des variables qualitatives à l'événement de référence "LOW = oui »......11
Tableau 6 : tests de normalité (Kolmogorov-Smirnov)......14
Tableau 7: tests de comparaisons non-paramétriques à deux échantillons de Wilcoxon......14

I Traitement des données :

Suites aux premières observations et manipulations effectuées sur les données, des recodages de variables et des modifications ont été jugées nécessaires pour une meilleure lisibilité des données.

L'unité du poids des variables LWT et BWT initialement en livre et gramme, sont converties en kilogramme.

Les variables qualitatives binaires sont recodées avec les références « oui », si celles-ci valent 1, et « non » si elles valent 0, pour la variable SMOKE, on donne à la valeur 1 « fumeur » et « non-fumeur » sinon. Concernant la variable RACE, on réassocie la référence « Blanche » (respectivement « Noire » et « Autre »), si cette variable vaut 1, (respectivement si elle vaut 2 ou 3).

On a décidé de traiter la variable quantitative AGE, en variable qualitative possédant 2 modalités. On a donc séparé les 189 femmes en 2 groupes catégorisés par 2 classes d'âges (14 à 23 ans et 24 à 45 ans), séparés par l'âge médian (23 ans) de l'échantillon afin de comparer une variable d'intérêt selon la classe d'âge donnée.

La variable ID portant sur les numéros des individus de l'échantillon est abandonnée car elle est inutile.

Dans le cadre d'une étude dans le domaine du médicale, on ne pourra pas supposer au départ que des variables quantitatives même à grands effectifs (n>30) que celles-ci ont une distribution Gaussienne. De plus ont ne cherchera pas à supprimer des valeurs qui semblent aberrantes car dans ce domaine, elles ont toutes une grande importance pour les enjeux de l'étude.

II Analyse univariée :

Une analyse descriptive simple a été réalisée sur les variables de l'ensemble de la population de l'étude. L'analyse bivariée et multivariée des facteurs associés au poids du bébé (BWT) et de la variable binaire associée à sa suffisance (LOW), a été conduite séparément.

Les caractéristiques physiques, pathologiques et de suivis des femmes interrogées par l'enquête sont résumés dans le tableau 1 concernant les variables quantitatives.

Pour les variables qualitatives, on les résume dans le tableau 2 en donnant les effectifs et proportions des classes de celles-ci.

Tableau 1 : statistiques descriptives des variables quantitatives.

Variables	Moyenne	Quartiles : inférieur	supérieur	Médiane	Ecart type	Minimum	Maximum
LWT	58.88	49.90	63.50	54.89	13.87	36.29	113.40
PTL	0.20	0.00	0.00	0.00	0.49	0.00	3.00
FVT	0.79	2.71	1.00	0.00	1.06	0.00	6.00
BWT	2.94	2.41	3.48	2.98	0.73	0.71	4.99

Figure 1 : boxplots de FVT et BWT.

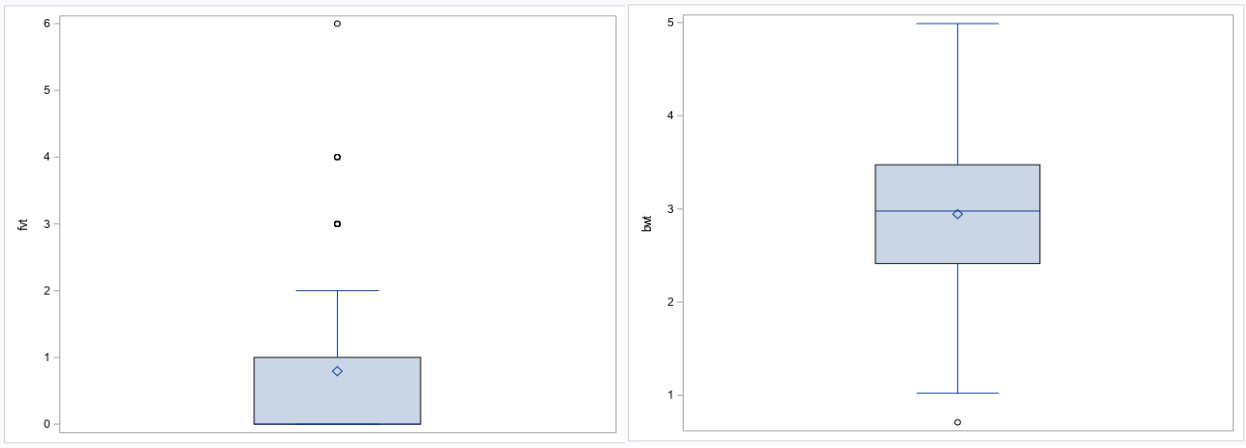
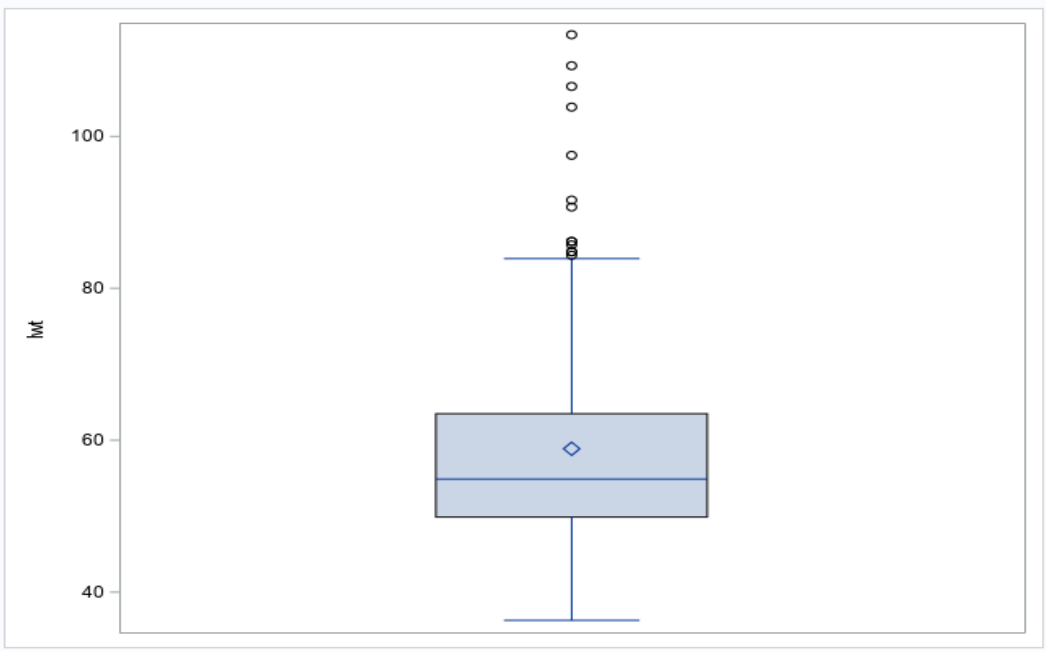


Figure 2 : boxplots de LWT et PTL.



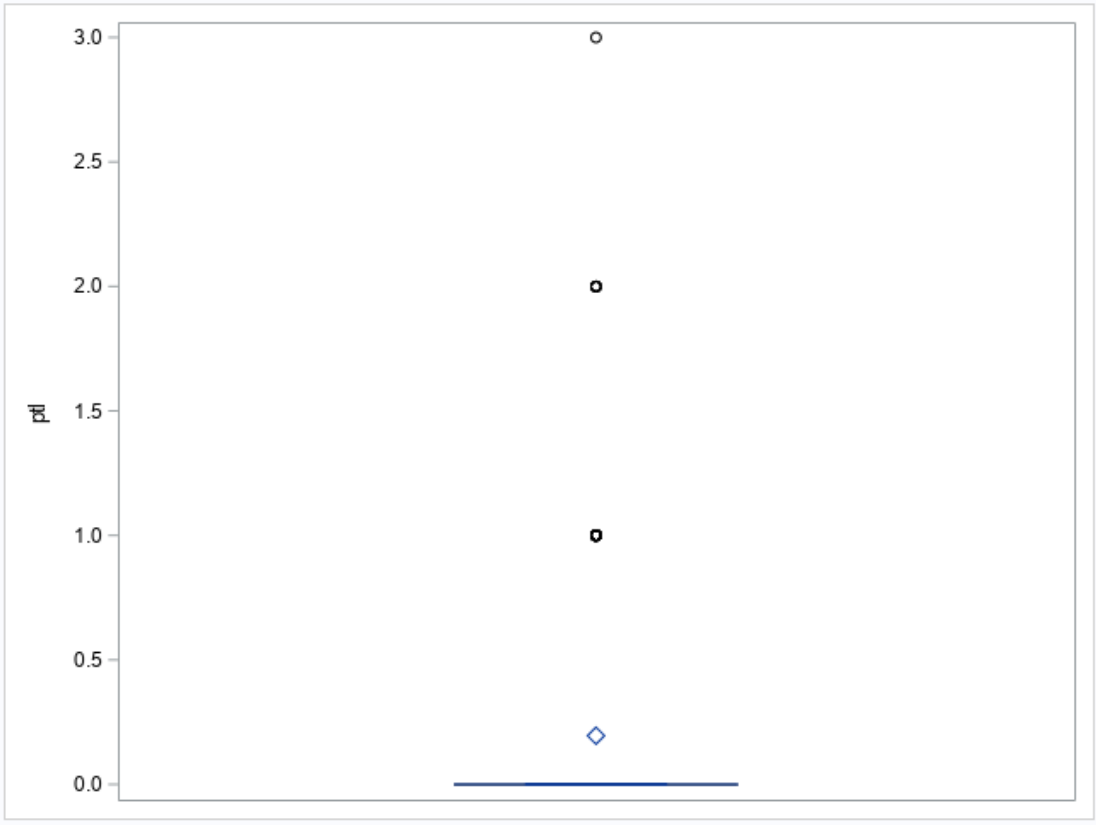
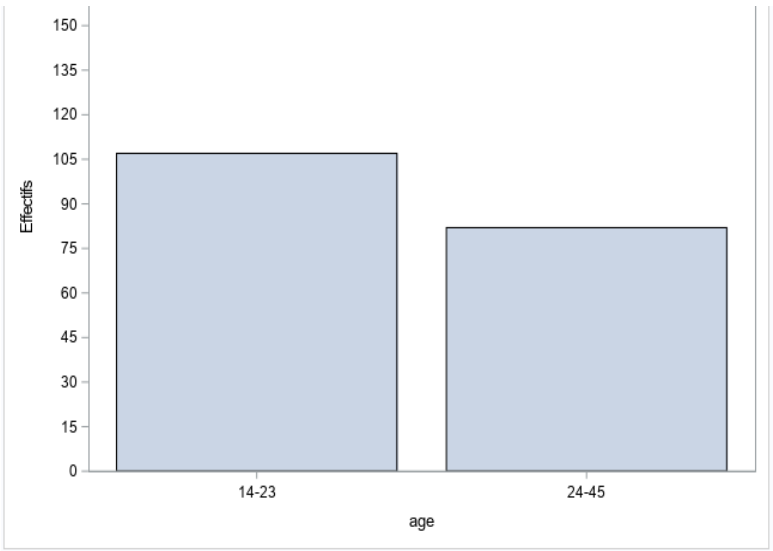


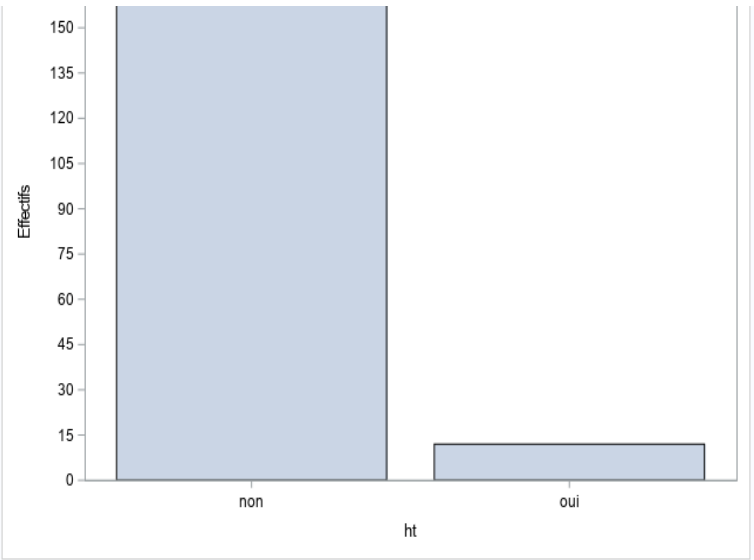
Tableau 2 : statistiques descriptives des variables qualitatives.

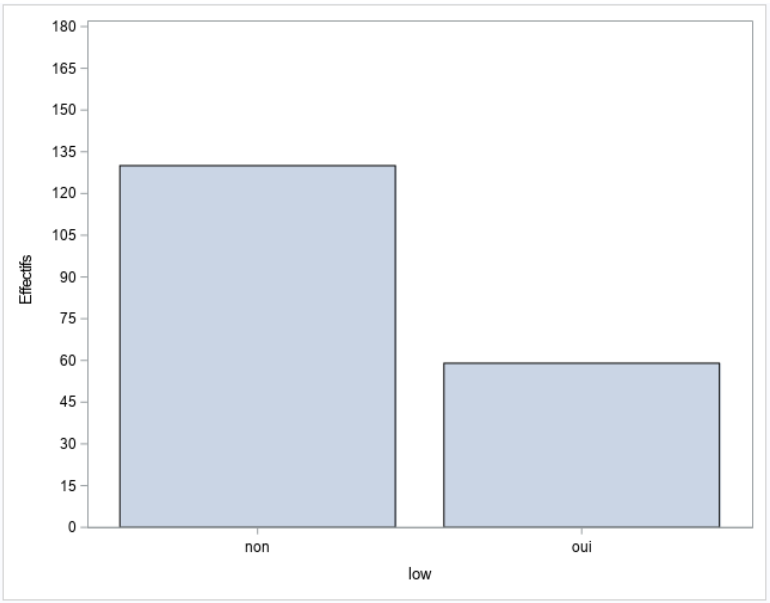
Variables	N *	% **
RACE : « Blanche » « Noire » « Autre »	 96 26 67	 50.79 13.76 35.45
SMOKE : « fumeur » « non-fumeur »	 74 115	 39.15 60.85
HT : « oui » « non »	 12 177	 6.35 93.65
UI : « oui » « non »	 28 161	 14.81 85.19
LOW : « oui » « non »	 59 130	 31.22 68.78
AGE : « 14-23 » « 24-45 »	 107 82	 56.61 43.39

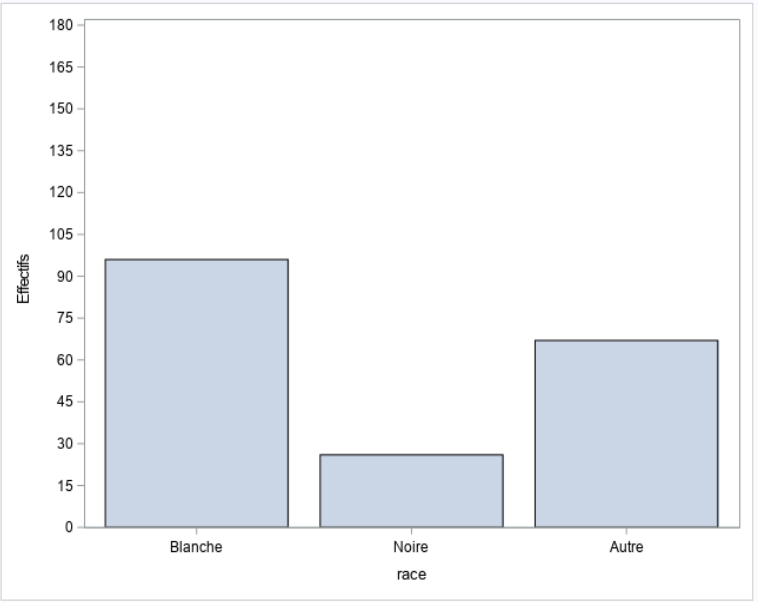
N : effectifs %** : pourcentages*

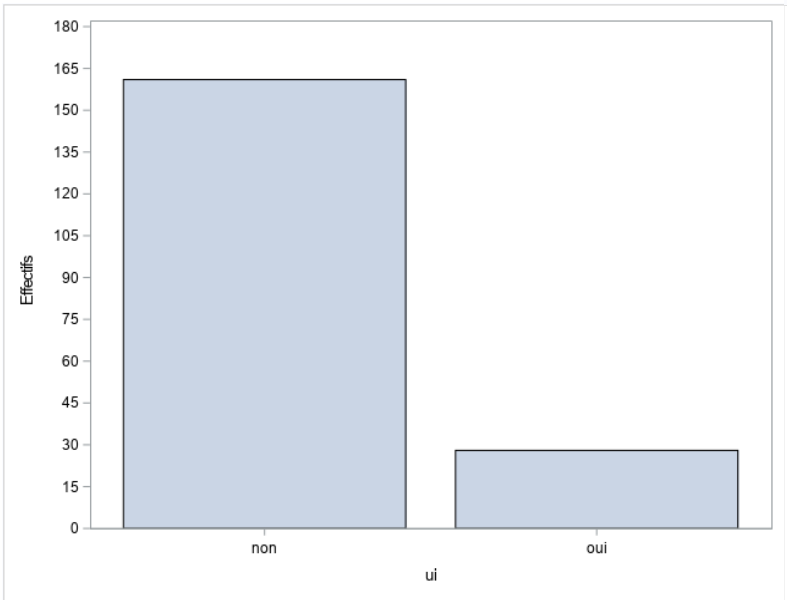
Figure 3 : diagrammes en bâtons des variables qualitatives.

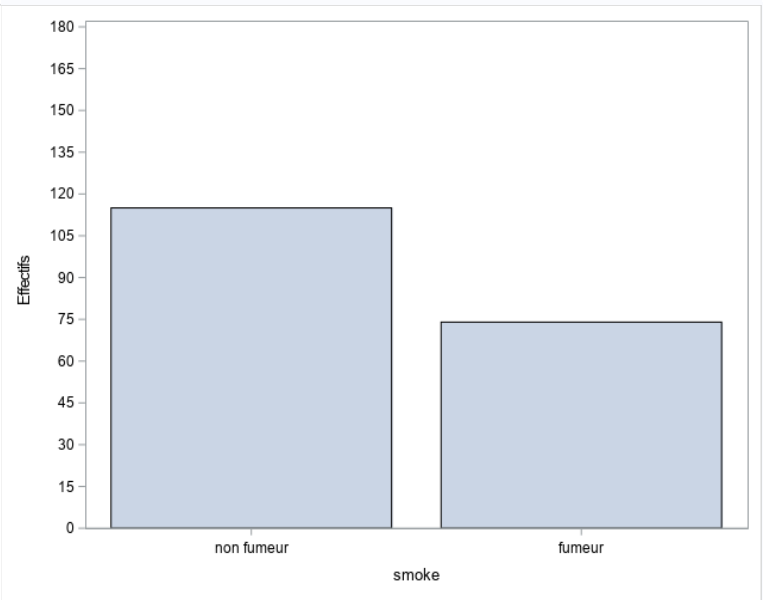












III Analyse bivariée :

Cette partie aura pour but de sélectionner les variables significatives sur les variables d'intérêts de l'étude (LOW et BWT), pour l'analyse multivariée.

Nous allons donc effectuer des tests d'indépendances de Fisher, entre la variable qualitative à expliquer désignant la suffisance du poids du bébé (LOW) avec les autres variables qualitatives étant des facteurs possibles de celle-ci.

Pour les variables quantitatives, nous allons regarder leurs distributions dans les 2 classes de la variable à expliquer LOW, pour voir quel test de comparaison de distributions est adapté. Si les deux échantillons suivent une loi normale, on utilisera le test paramétrique d'égalité de moyennes de Student, sinon le test non-paramétrique de Wilcoxon.

1) Variables qualitatives :

a) Tests d'indépendance de Fisher.

Pour la variable HT, un test exact de Fisher est réalisé car elle possède des cellules avec un effectif théorique inférieur à 5.

Tableau 3 : Résultats du test exact de Fisher.

Variables	P-value
HT	0.052

Tableau 4 : Résultats du test de Fisher.

Variables	P-value
SMOKE	0.040
UI	0.035
RACE	0.082
AGE	0.728

On peut supposer d'après ces résultats que les facteurs à modalités ayant une incidence sur le poids de naissance du bébé, sont la présence d'irritabilité utérine et la consommation de tabac durant la grossesse. Nous allons donc garder ces variables dans l'analyse multivariée, car les p-value des tests d'indépendance entre LOW et ces variables sont inférieur à 5%.

On va s'intéresser aux odds ratios à l'événement LOW = « oui » (bébé en sous poids) par rapport aux modalités désignant la présence des facteurs retenus, pour voir s'ils sont cohérents avec nos hypothèses en ajoutant HT pour qui la p-value du test d'indépendance est très proche de 5%.

b) Calcul des odds ratio.

L'odds ratio (OR), également appelé rapport des cotes, est une mesure statistique exprimant le degré de dépendance entre des variables aléatoires qualitatives.

Si la probabilité qu'un événement arrive dans le groupe A est *p*, et *q* dans le groupe B, le rapport des cotes est :

$$\frac{p/(1-p)}{q/(1-q)} = \frac{p(1-q)}{q(1-p)}$$

Interprétations des OR d'un événement :

S'il est proche de 1, l'évènement est indépendant du groupe.

Lorsque l'OR est > 1, cela signifie que l'événement est plus fréquent dans le groupe A que dans le groupe B. Lorsque l'OR est <1, l'événement est moins fréquent dans le groupe A que dans le groupe B.

Ici on va calculer les OR pour l'événement : LOW = « oui » (le bébé est en sous poids) pour les modalités désignant la présence des variables qualitatives significatives sur LOW et HT.

Tableau 5 : rapport des cotes des variables qualitatives à l'événement de référence "LOW = oui ».

Variables	OR
HT = « oui »	3.365
SMOKE = « fumeur »	2.022
UI = « oui »	2.578

Les OR de LOW = « oui » associés aux modalités des variables ci-dessus, sont bien supérieurs à 1 donc on peut supposer que les femmes possédants ces symptômes ou traits suivants, présentent un risque élevé d'avoir un bébé en sous poids. HT sera gardée dans l'analyse multivariée car pour la modalité HT = « oui », l'OR est le plus élevé.

c) Représentations graphiques de la distribution du poids de naissance des bébés en fonction des facteurs à modalités a priori significatif.

Dans les diagrammes en boîtes suivants, on voit que selon la classe des variables qualitatives significatives, que la distribution de la variable quantitative du poids du bébé (BWT), présente des différences de niveaux de quartiles inférieurs, de médianes et de moyennes qui ne semblent pas négligeable :

Figure 4 : distribution de BWT en fonction de HT.

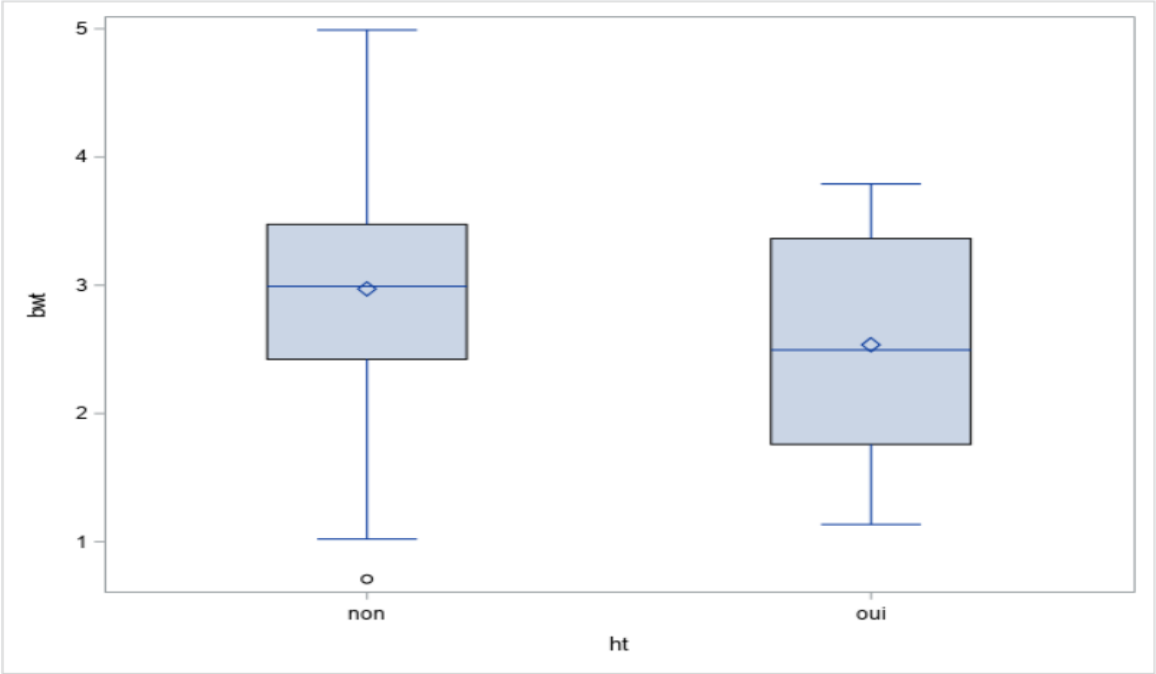


Figure 5 : distribution de BWT en fonction de UI.

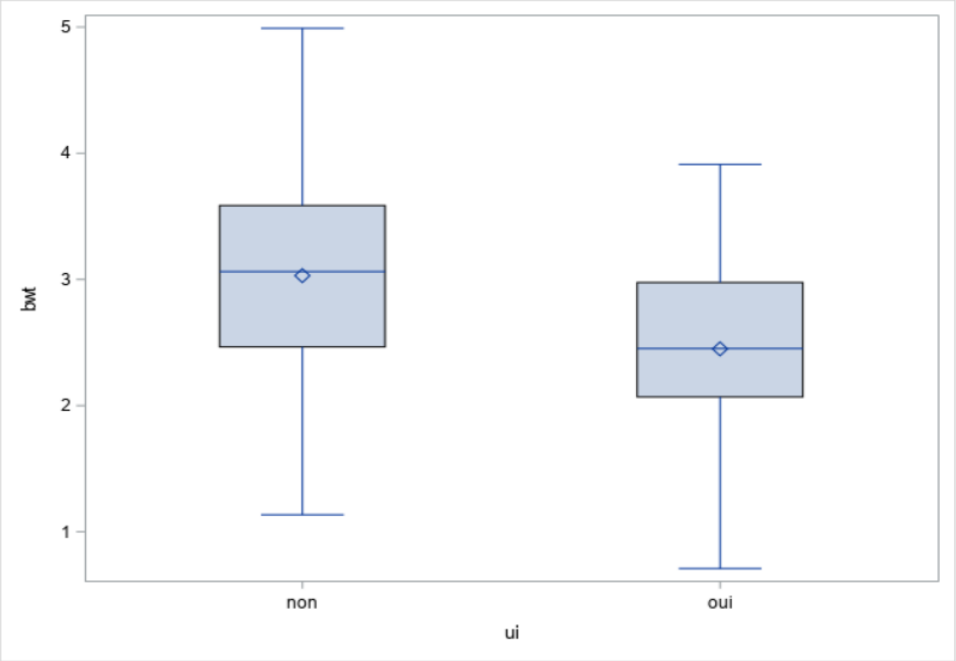
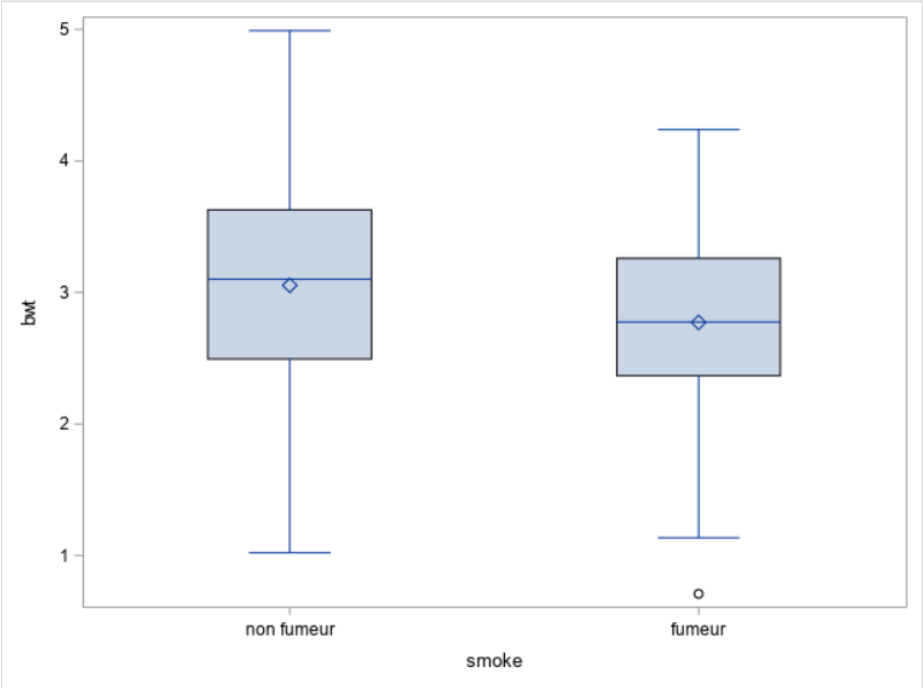


Figure 6 : distribution de BWT en fonction de SMOKE.



2) Variables quantitatives :

Pour sélectionner les variables quantitatives explicatives que nous allons utiliser dans l'analyse multivariée, nous allons faire des tests de comparaisons de moyennes de celles-ci selon la classe de LOW, donc avec 2 échantillons indépendants (un avec LOW = « oui », l'autre avec LOW = « non »).

Tout d'abord nous allons tester la normalité des distributions des 2 échantillons indépendants des variables quantitatives explicatives (selon la classe de LOW) pour déterminer les tests de comparaisons appropriés de ces derniers.

a) Tests de normalité à 2 échantillons d'une variable.

Tableau 6 : tests de normalité (Kolmogorov-Smirnov).

Variables	P-value (LOW= « oui »)	P-value (LOW= « non »)
LWT	<0.01	<0.01
PTL	<0.01	<0.01
FVT	<0.01	<0.01

Aucune variable n'a ses 2 échantillons ensembles qui suivent la loi normale (on accepte l'hypothèse de normalité quand la p-value > 0.05), donc le test paramétrique de Student pour comparer l'égalité des moyennes de 2 échantillons ne convient pas, on va utiliser le test non-paramétrique de comparaison de Wilcoxon pour comparer les distributions entre les deux échantillons.

b) Tests non-paramétriques de comparaison.

Tableau 7: tests de comparaisons non-paramétriques à deux échantillons de Wilcoxon.

Variables	P-value
LWT	0.014
PTL	<0.001
FVT	0.240

On constate donc que les distributions selon les modalités de LOW diffèrent pour LWT et PTL (p-value < 0.05).

Voici dans le tableau 8, les coefficients de coefficients de corrélation de Spearman. Ils estiment à quel point la relation entre deux variables peut être décrite par une fonction monotone, les valeurs de la dernière colonne sont celles qui nous intéressent. Plus le coefficient en valeur absolue est proche de 1, plus la corrélation entre les 2 variables est forte. S'il est positif, les variables auront tendance à évoluer dans le même sens, et dans le sens contraire sinon.

Coefficients de corrélation de Spearman, N = 189 Proba > |r| sous H0: Rho=0					
	age	lwt	ptl	fvt	bwt
age	1.00000	0.18606 0.0104	0.11850 0.1044	0.23417 0.0012	0.06107 0.4038
lwt	0.18606 0.0104	1.00000	-0.11086 0.1289	0.08902 0.2232	0.24831 0.0006
ptl	0.11850 0.1044	-0.11086 0.1289	1.00000	-0.01432 0.8449	-0.20393 0.0049
fvt	0.23417 0.0012	0.08902 0.2232	-0.01432 0.8449	1.00000	0.07010 0.3378
bwt	0.06107 0.4038	0.24831 0.0006	-0.20393 0.0049	0.07010 0.3378	1.00000

Annexe 2 : table des coefficients de Spearman entre les variables quantitatives.

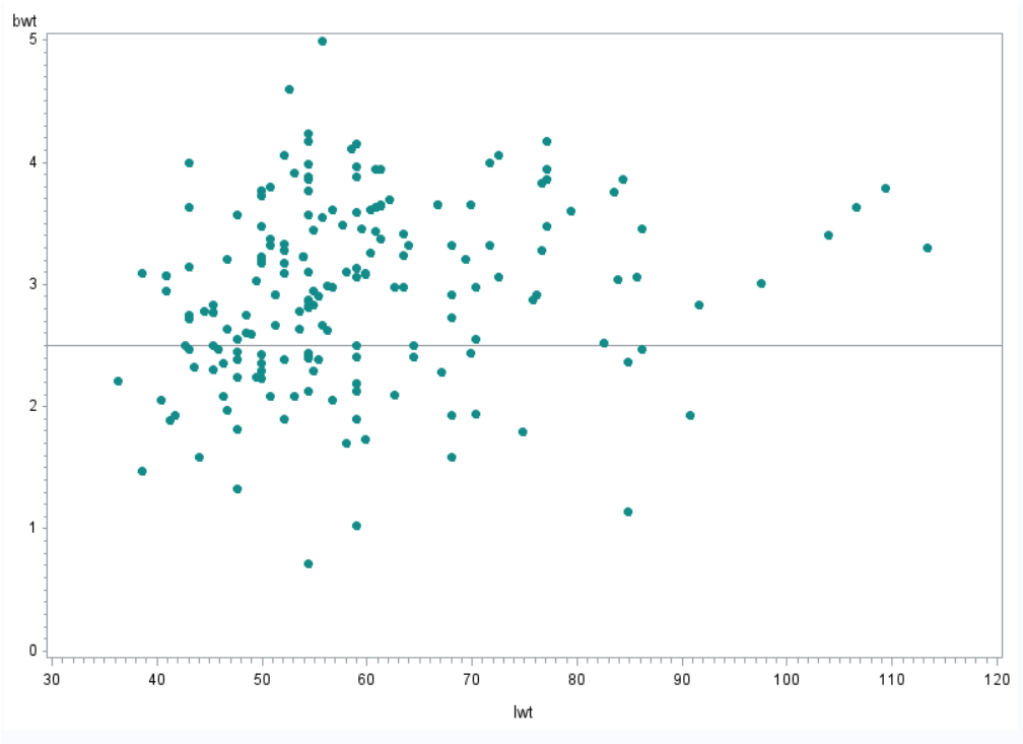
A noter qu'on a rajouter la variable AGE avec son format initial (quantitative) pour voir si le résultat du test de Fisher entre LOW et AGE (avec modalités) est cohérent avec la valeur du coefficient calculé.

La corrélation entre FVT (respectivement AGE) et BWT (le poids du bébé) est très faible : environ 6% (respectivement environ 7%).

En visualisant les nuages de points du poids du bébé en kilogrammes (BWT) en fonction de PTL et LWT, en y ajoutant la ligne horizontale désignant seuil de suffisance du poids du bébé (2,5 kg), il semblerait que plus le poids de la mère au dernier cycle menstruel est faible, plus il y a de bébé à la naissance dont le poids est en dessous du seuil de suffisance.

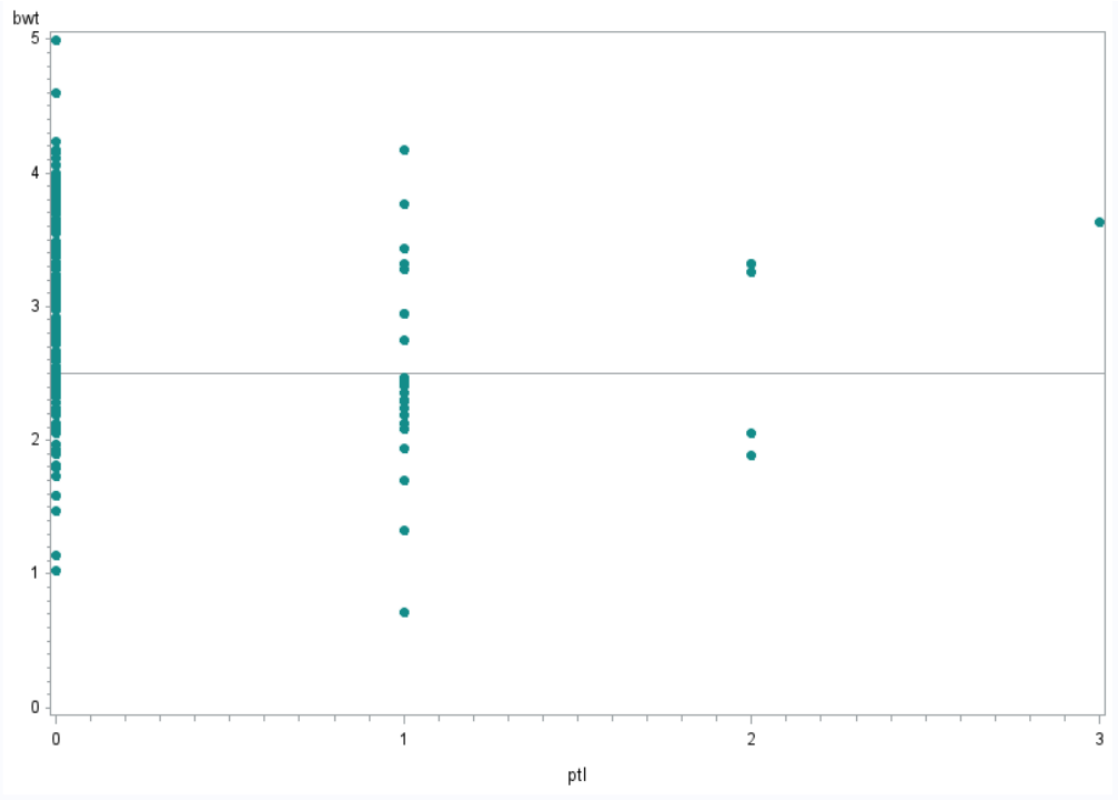
En comparant les nombres de points en dessous et au-dessus de la ligne horizontale, on peut penser que plus y a eu des antécédents de prématurités pour la mère, plus l'effectif du nombre de bébé en sous poids est proportionnellement grand par rapport à l'effectif du nombre de mère avec des bébé au-dessus du seuil de suffisance, ce nuage de points est difficile à lire. Nous verrons ces propriétés plus en détails dans l'analyse multivariée.

Figure 7 : nuage de points de BWT en fonction de LWT.



.

Figure 8 : nuage de points de BWT en fonction de PTL.



IV Analyse multivariée :

1) Régression logistique multiple.

Pour déterminer laquelle des variables LWT et PTL a l'influence la plus forte sur LOW, on décide d'ajuster un modèle de régression logistique aux données, avec sélection de variables par la méthode « backward » appuyée par le test de Wald.

Ici, c'est la modalité 1 (« oui ») de la variable de réponse LOW qui est la valeur dont on souhaite modéliser la probabilité en fonction de PTL et LWT.

Tableau 8 : résultats du test de Wald.

Variables	P-value
PTL*LWT	0.427
LWT	0.046
PTL	0.026

On constate que le poids de la mère lors du dernier cycle menstruel et le nombre d'antécédent de prématurité est retenu comme facteurs explicatifs par la méthode de sélection de variables.

L'effet conjoint PTL*LWT a été supprimé (p-value>0.05) et d'après le test, la variable ayant le plus d'influence sur LOW est PTL car il a une p-value inférieure à celle de LWT.

Figure 9 : probabilité d'avoir LOW = "oui" en fonction de LWT.

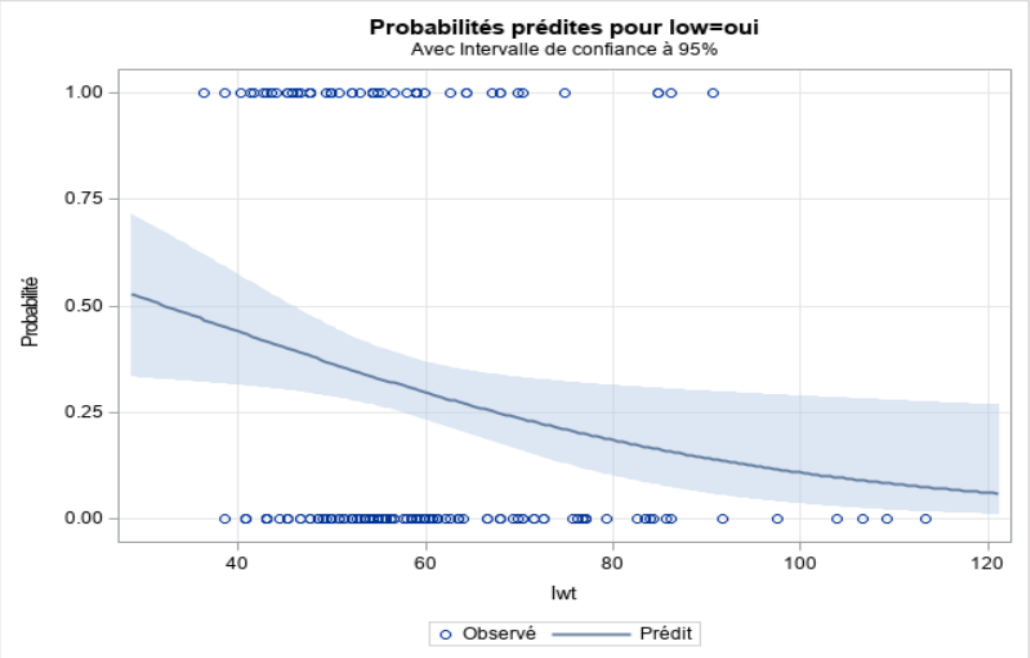
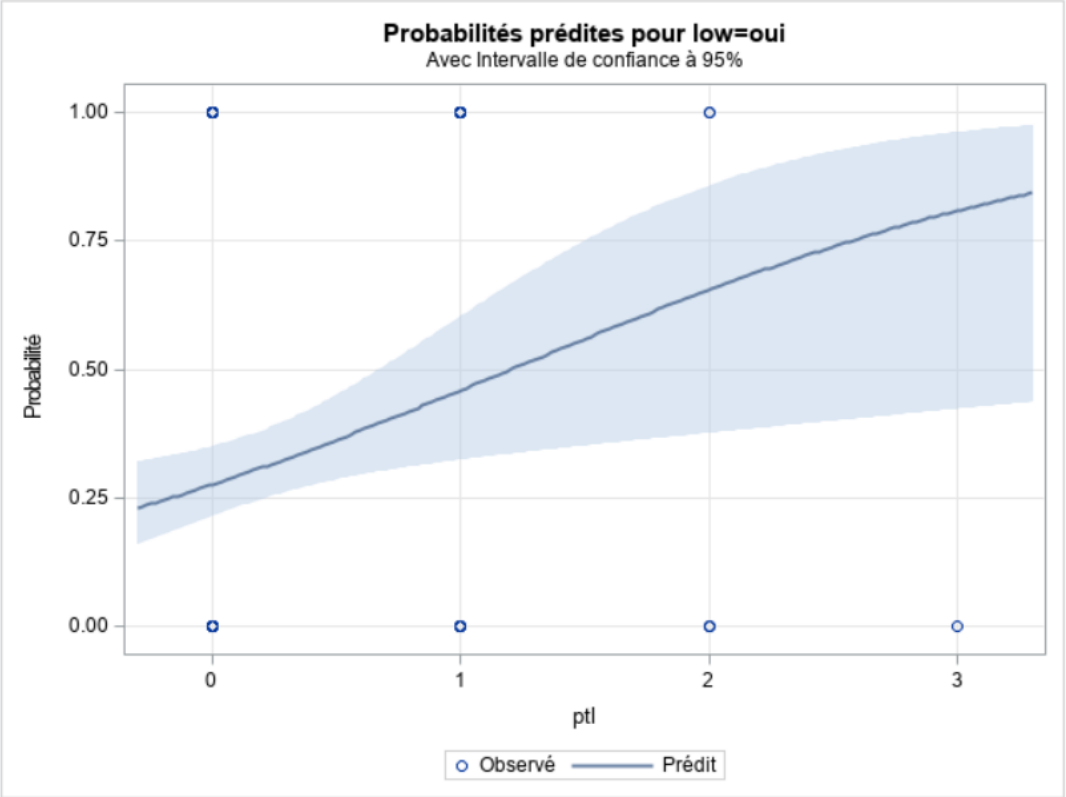


Figure 10 : probabilité d'avoir LOW = "oui" en fonction de PTL.



On constate que d'après ces modèles de régressions logistiques (LOW en fonction de LWT et LOW en fonction de PTL) que la probabilité d'avoir un bébé en sous poids, augmente avec le nombre d'antécédents de prématurités et qu'elle diminue plus le poids de la mère au dernier cycle menstruel est élevée.

2) Analyse en correspondance multiples :

L'analyse des correspondances multiples (ACM) est une méthode factorielle adaptée aux tableaux dans lesquels un ensemble d'individus en lignes est décrit par un ensemble de variables qualitatives et ses modalités en colonnes (tableau disjonctif complet).

Le principe consiste à résumer l'information contenu de cet ensemble de variables associé aux individus, afin de faciliter l'interprétation des corrélations existantes entre ces différentes variables grâce au tableau disjonctif complet. Ce tableau donne aussi les profils de réponses des individus. On cherche à savoir quelles sont les modalités corrélées entre elles, et le(s) profil(s) de réponses d'individus d'une modalité d'une variable d'intérêt.

Cela est représenté graphiquement dans un plan factoriel.

Ainsi, il est possible de déterminer des proximités entre des modalités de variables différentes ou entre individus pour en tirer des enseignements :

-Les individus qui ont beaucoup de modalités en commun sont aussi proches que possible.

-Les individus qui ont peu ou aucunes modalités en commun sont aussi séparés que possible.

-Deux modalités sont proches si elles ont souvent été prises ensemble.

Comme variables qualitatives actives, nous avons introduit toutes les variables significatives pouvant permettre d'expliquer l'insuffisance de poids d'un bébé (HT, UI, SMOKE). Ces variables comportent au total 6 modalités actives. La variable LOW est mis supplémentaire car elle est celle à expliquer.

On décide de retenir 2 axes, car leurs pourcentages d'inerties cumulées (71%) est suffisant pour une interprétation de cette ACM à 3 variables actives.

Décomposition de l'inertie et du Khi-2

Valeur singulière	Inertie principale	Khi-2	Pourcentage	Pourcentage cumulé
0.61092	0.37322	213.852	37.32	37.32
0.58067	0.33718	193.197	33.72	71.04
0.53814	0.28960	165.936	28.96	100.00
	1.00000	572.985	100.00	

Degrés de liberté = 25

Annexe 3 : décomposition de l'inertie et du khi-2.

Interprétation des résultats :

Pour chaque axe, le pourcentage d'inertie théorique moyen expliqué par chaque modalité est de 16.67% (100%/6).

Seules les modalités dont la contribution est élevée sont à considérer pour l'interprétation d'un axe c'est-à-dire celles dont la contribution est supérieure à 16.67%.

Les modalités de la variable LOW mises en éléments supplémentaires serviront également à caractériser les axes et ne contribuent pas à l'inertie. Par exemple, si la modalité « oui » de la variable SMOKE se retrouvait du côté positif de l'axe 1 (coordonnée positive), on pourrait s'attendre à ce que les individus possédant les modalités actives biens représentées du côté positif de l'axe 1 (contribution est supérieure à 16.67%), aient plus de chance d'avoir comme modalité « oui » de la variable LOW. Les modalités actives à considérer pour l'interprétation des 2 axes, sont consignées dans le tableau suivant.

Tableau 9 : caractérisation des axes factoriels.

côté positif	côté négatif
Axe 1	
	UI = « oui »
HT = « oui »	
Axe 2	
HT = « oui »	
	SMOKE = « non-fumeur »
SMOKE = « fumeur »	

Par soucis de lecture des modalités dans l'ACM, les modalités des variables LOW, HT, UI sont recodées de la forme : « variable modalité ». Les carrés du cosinus des points selon les axes des variables dont la contribution partielle à l'inertie des points des colonnes sont très significatifs (on considère arbitrairement, supérieur à 10 %), ces points sont bien représentés sur l'axe associé donc fiables pour l'interprétation.

Contributions partielles à l'inertie des points des colonnes

	Dim1	Dim2
ht non	0.0245	0.0153
ht oui	0.3610	0.2264
fumeur	0.0568	0.4595
non-fumeur	0.0365	0.2957
ui non	0.0772	0.0005
ui oui	0.4440	0.0026

Carrés du cosinus pour les points des colonnes

	Dim1	Dim2
ht non	0.4316	0.2445
ht oui	0.4316	0.2445
fumeur	0.1045	0.7640
non-fumeur	0.1045	0.7640
ui non	0.5836	0.0031
ui oui	0.5836	0.0031

Coordonnées des colonnes

	Dim1	Dim2
ht non	-0.1711	-0.1287
ht oui	2.5231	1.8990
fumeur	-0.4029	1.0896
non-fumeur	0.2593	-0.7011
ui non	0.3186	-0.0231
ui oui	-1.8319	0.1329

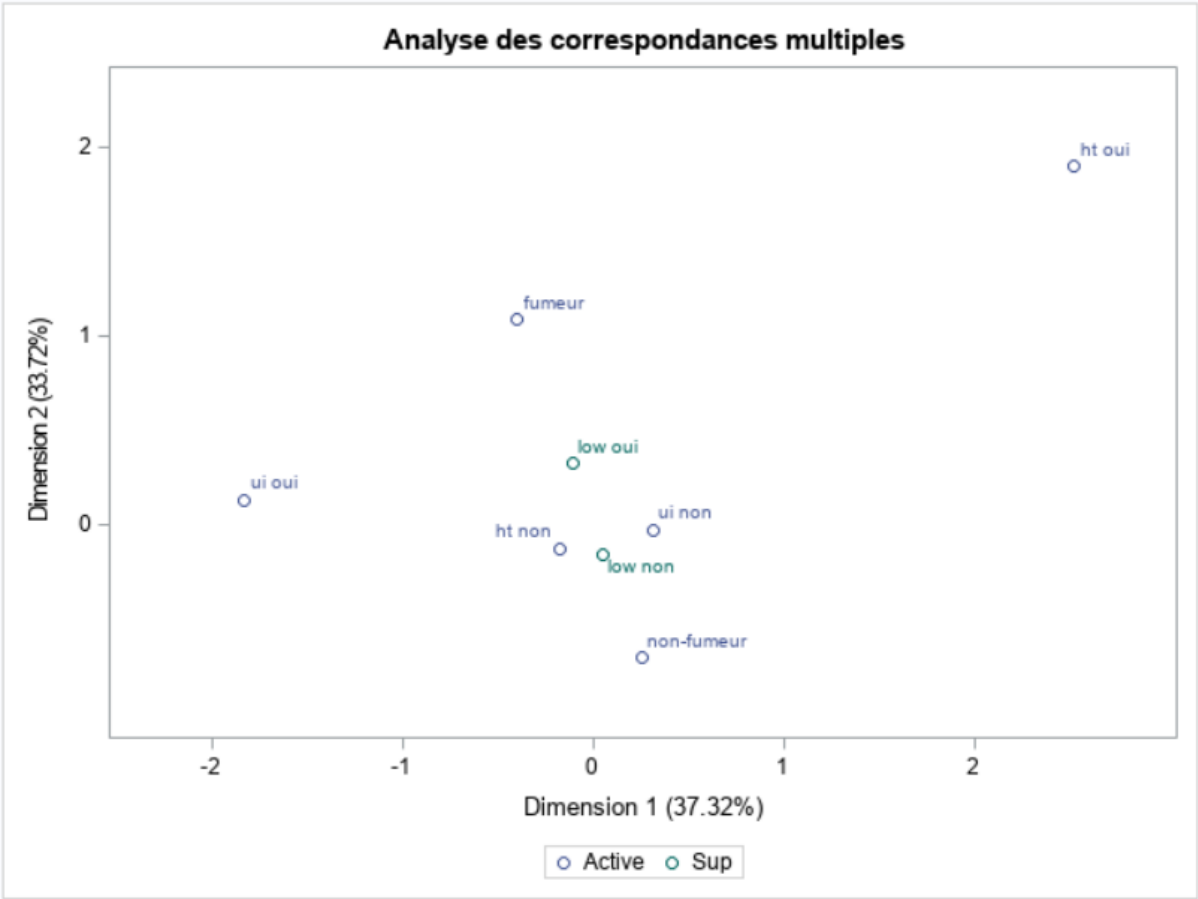
Coordonnées des colonnes supplémentaires

	Dim1	Dim2
low non	0.0489	-0.1504
low oui	-0.1077	0.3313

Annexe 4 : contributions à l'inertie et carrés du cosinus des variables actives de l'ACM, et coordonnées de toutes les variables.

La modalité LOW = « oui » se trouve du côté négatif sur l'axe 1 et du côté positif de l'axe 2 ». A l'inverse la modalité LOW = « non », se trouve du côté négatif sur l'axe 2 et du côté positif de l'axe 1 ».

Figure 11 : représentation graphique de l'ACM.



On constate donc que les femmes susceptibles d'avoir des bébés de poids inférieur à 2,5 kg sont celles qui ont eu des antécédents d'hypertension (le cosinus au carré de HT = « oui » est plus grand sur l'axe 1). Les femmes ne fumant pas durant la grossesse, sont favorable au fait d'avoir un bébé avec un poids suffisant. Ce qui confirme nos hypothèses de l'analyse bivariée.

Conclusion :

Les résultats de l'étude des facteurs du poids de naissance des bébés à mené à retenir comme facteurs explicatifs à un faible poids, le tabagisme durant la grossesse, la présence d'irritabilité utérine, le nombre d'antécédents de prématurité et le poids de la mère lors du dernier cycle menstruel. Quand celui-ci devient faible, la probabilité de chance que le poids du bébé soit insuffisant qui va naître augmente, cette probabilité augmente aussi, plus la femme possède des antécédents de prématurité. Les plus simples recommandations aux femmes pour réduire les chances d'avoir un bébé avec un poids en dessous du seuil de suffisance (2,5 kg)

qu'on pourrait fournir sont d'éviter le tabac, et de manger insuffisamment pendant la grossesse. Dans le domaine de la médecine, il est difficile de déjouer les pathologies apparues pour causes héréditaires, on peut néanmoins si on est conscient leur présence au moins freiner leur impact sur l'organisme.

Comme extension de l'étude, on pourrait étendre le nombre d'individus en ajoutant leurs pays d'origine. Pour voir comment le poids de naissance du bébé, pourrait être influé par le mode de vie d'un pays à l'autre, le niveau de richesse, et si le temps au travail (par rapport à celui consacré aux distractions) pouvant entraîner un stress des individus est néfaste au bon développement du bébé pendant la grossesse.

Références :

Jeu de données :

http://www.biostatisticien.eu/springeR/jeuxDonnees2.html

http://www.jybaudot.fr/Analdonnees/acm.html

https://fr.wikipedia.org/wiki/Odds_ratio

Annexes :

Annexe 1 : liste des variables de l'étude.

Annexe 2 : table des coefficients de Spearman entre les variables quantitatives.

Annexe 3 : décomposition de l'inertie et du khi-2.

Annexe 4 : contributions à l'inertie et carrés du cosinus des variables actives de l'ACM, et coordonnées de toutes les variables.

Annexe 5 : vue partielle des données avant traitements.

Annexe 6 : vue partielle des données après traitements.

Codes sous SAS :

Analyse univariée :

Dans les instructions `%let var = (groupe de variable en commentaire)`

On y place la variable qu'on souhaite utilisé, et &var prendra sa valeur.

```
/*traitements et lecture des données*/

data poids;
infile 'C:\Users\Ascensio\Downloads\TDSAS\Poids_naissance.csv' dsd missover
dlm=';' firstobs=2;
input id age lwt race smoke ptl ht ui fvt bwt low;
run;
PROC FORMAT;
VALUE race
1="Blanche"
2 ="Noire"
3="Autre";

VALUE smoke
0="non-fumeur"
1="fumeur";

VALUE ht
1="oui"
0="non";
```

```
VALUE ui
1="oui"
0="non";

VALUE low
1="oui"
0="non";

VALUE classAge
14-<24 = "14-23"
24-45 = "24-45";
run;

data poidsChanged;
infile 'C:\Users\Ascensio\Downloads\TDSAS\Poids_naissance.csv' dsd missover
dlm=';' firstobs=2;
input id age lwt race smoke ptl ht ui fvt bwt low;
lwt=lwt/2.2046;
bwt=bwt/1000;
drop id;
FORMAT age $classAge. race $race. smoke $smoke. ht $ht. ui $ui. low $low.;
run;
```

Obs.	id	age	lwt	race	smoke	ptl	ht	ui	fvt	bwt	low
1	85	19	182	2	0	0	0	1	0	2523	0
2	86	33	155	3	0	0	0	0	3	2551	0
3	87	20	105	1	1	0	0	0	1	2557	0
4	88	21	108	1	1	0	0	1	2	2594	0
5	89	18	107	1	1	0	0	1	0	2600	0
6	91	21	124	3	0	0	0	0	0	2622	0
7	92	22	118	1	0	0	0	0	1	2637	0
8	93	17	103	3	0	0	0	0	1	2637	0
9	94	29	123	1	1	0	0	0	1	2663	0
10	95	26	113	1	1	0	0	0	0	2665	0
11	96	19	95	3	0	0	0	0	0	2722	0
12	97	19	150	3	0	0	0	0	1	2733	0
13	98	22	95	3	0	0	1	0	0	2750	0
14	99	30	107	3	0	1	0	1	2	2750	0
15	100	18	100	1	1	0	0	0	0	2769	0
16	101	18	100	1	1	0	0	0	0	2769	0
17	102	15	98	2	0	0	0	0	0	2778	0
18	103	25	118	1	1	0	0	0	3	2782	0
19	104	20	120	3	0	0	0	1	0	2807	0
20	105	28	120	1	1	0	0	0	1	2821	0

Annexe 5 : vue partielle des données avant traitements.

Obs.	age	lwt	race	smoke	ptl	ht	ui	fvt	bwt	low
1	14-23	82.555	Noire	non-fumeur	0	ht oui	ui non	0	2.523	low non
2	24-45	70.308	Autre	non-fumeur	0	ht oui	ui oui	3	2.551	low non
3	14-23	47.628	Blanche	fumeur	0	ht oui	ui oui	1	2.557	low non
4	14-23	48.988	Blanche	fumeur	0	ht oui	ui non	2	2.594	low non
5	14-23	48.535	Blanche	fumeur	0	ht oui	ui non	0	2.600	low non
6	14-23	56.246	Autre	non-fumeur	0	ht oui	ui oui	0	2.622	low non
7	14-23	53.524	Blanche	non-fumeur	0	ht oui	ui oui	1	2.637	low non
8	14-23	46.720	Autre	non-fumeur	0	ht oui	ui oui	1	2.637	low non
9	24-45	55.792	Blanche	fumeur	0	ht oui	ui oui	1	2.663	low non
10	24-45	51.256	Blanche	fumeur	0	ht oui	ui oui	0	2.665	low non
11	14-23	43.092	Autre	non-fumeur	0	ht oui	ui oui	0	2.722	low non
12	14-23	68.040	Autre	non-fumeur	0	ht oui	ui oui	1	2.733	low non
13	14-23	43.092	Autre	non-fumeur	0	ht non	ui oui	0	2.750	low non
14	24-45	48.535	Autre	non-fumeur	1	ht oui	ui non	2	2.750	low non
15	14-23	45.360	Blanche	fumeur	0	ht oui	ui oui	0	2.769	low non
16	14-23	45.360	Blanche	fumeur	0	ht oui	ui oui	0	2.769	low non
17	14-23	44.453	Noire	non-fumeur	0	ht oui	ui oui	0	2.778	low non
18	24-45	53.524	Blanche	fumeur	0	ht oui	ui oui	3	2.782	low non
19	14-23	54.432	Autre	non-fumeur	0	ht oui	ui non	0	2.807	low non
20	24-45	54.432	Blanche	fumeur	0	ht oui	ui oui	1	2.821	low non

Annexe 6 : vue partielle des données après traitements.

Statistiques descriptives :

```
proc means data = poidsChanged Q1 median Q3 qrange mean std min max maxdec
= 3;
var age lwt ptl fvt bwt; /* pour récupérer l'âge median */
run;

/*boxplots des variables quantitatives */

%let var = ptl/* bwt lwt fvt   */;
proc sgplot data = poids;
    vbox &var;
run;

/*pourcentages des variables qualitatives*/

%let var = low/*age ht smoke race ui*/;
proc freq data = poidsChanged;
table age;
run;
/* diagrammes en bâtons des variables qualitatives */
```

```
%let var = smoke/*ht age ui race low*/;
proc sgplot data = poidsChanged;
yaxis label= "Effectifs"
values =(0 to 189 by 15);
vbar &var ;
run;
```

Analyse bivariée :

```
/* tests de khi-2 de low avec chaque autre variable qualitative */
%let var =   age/*ht race smoke ui */;
proc freq data = poidsChanged;
table low*&var/chisq nocum norow;
run;

/*tests de normalité des variables quantitatives à 2 échantillons classés
par low*/

%let var = age /*lwt fvt ptl */;

proc univariate normaltest plot data = poids;
      var &var;
      class low;
      histogram / normal kernel;
      run;

/*tests de comparaisons des distributions de variables quantitatives à 2
échantillons classés par low*/


proc npar1way data= poids wilcoxon;
class low;
var &var;
run;



/*corrélations de spearman entre les variables quantitatives*/

proc corr data = poids  spearman ;
var age lwt ptl fvt bwt;

run;

/*nuages de points de bwt en fonction de lwt puis bwt en fonction de ptl*/

proc gplot data = poidsChanged;
plot bwt*lwt / OVERLAY VREF = 2.5;
run;



proc gplot data = poidsChanged;
plot bwt*ptl / OVERLAY VREF = 2.5;
run;
```

```
/*boxplots du poids bwt par rapport aux modalités de chaque variable
qualitative */

proc sgplot data = poidsChanged;
      vbox bwt/ category = &var;
run;


/*calculs des OR de LOW = « low oui » associés aux modalités de variables
qualitatives */

proc logistic data = poidsChanged;
class
      low (param=ref ref="low oui")
      smoke (param=ref ref="fumeur");
model low = smoke;
run;

proc logistic data = poidsChanged;
class
      low (param=ref ref="low oui")
      ui (param=ref ref="oui");
model low = ui;
run;

proc logistic data = poidsChanged;
class
      low (param=ref ref="low oui")
      ht (param=ref ref="oui");
model low = ht;
run;
```

Analyse multivariée :

```
/*régression logistique avec sélection de variables de low en fonction de
ptl et lwt*/

proc logistic data = poidsChanged desc;
model low = ptl lwt ptl*lwt / selection = backward;
run;


proc logistic data = poidsChanged desc plots = effect;
model low = ptl;
run;


proc logistic data = poidsChanged desc plots = effect;
model low = lwt;
run;

/*ACM*/

proc corresp data = poidsChanged mca dimens = 2 NOROW=PRINT ;
        tables ht smoke ui low ;
            sup low;
run;
```

Codes sous R :

```
"importation et traitements des données"

poids<- read.csv("~/Poids_naissance.csv", sep=";")
summary(poids)

"mise en place de classes pour la variable AGE"

for (i in 1:189) {
 if (poids$AGE[i] < 24)
 {poids$AGE [i] <- "14-23" }
 else  {poids$AGE [i] <- "24-45"}
}

"recodages des variables"

for (i in 1:189) {
 poids$LWT[i]<-poids$LWT[i]/2.2046
 poids$BWT[i]<-poids$BWT[i]/1000

 if (poids$RACE[i] == 1)
 {poids$RACE [i] <- "Blanche"}
 if (poids$RACE[i] == 2)
 {poids$RACE [i] <- "Noire"}
 if (poids$RACE[i] == 3)
 {poids$RACE [i] <- "Autre"}

 if (poids$SMOKE[i] == 1)
 {poids$SMOKE [i] <- "fumeur" }
 else {poids$SMOKE [i] <- "non-fumeur" }

 if (poids$HT[i] == 1)
 {poids$HT[i] <- "ht oui"}
 else  {poids$HT[i] <- "ht non"}

 if (poids$UI[i] == 1)
 {poids$UI [i] <- "ui oui"}
 else  {poids$UI [i] <- "ui non"}    }
```

```
"recodages de LOW"
for (i in 1:189) {
  if (poids$LOW[i] == 1)
  {poids$LOW [i] <- "low oui"}
  else  {poids$LOW [i] <- "low non"}
}
"suppression de ID"
poids<- poids[,-1]
"boxplots des variables quantitatives"
boxplot(poids$AGE)
boxplot(poids$LWT)
boxplot(poids$PTL)
boxplot(poids$BWT)
boxplot(poids$FVT)

"effectifs, pourcentages, diagrammes en barres des fréquences des classes des variables qualitatives"
table(poids$AGE)
barplot(table(poids$AGE)/189,col="blue", ylab="fréquence")

table(poids$SMOKE)
barplot(table(poids$SMOKE)/189,col="blue", ylab="fréquence")

table(poids$RACE)
barplot(table(poids$RACE)/189,col="blue", ylab="fréquence")

table(poids$HT)
barplot(table(poids$HT)/189,col="blue", ylab="fréquence")

table(poids$UI)
barplot(table(poids$UI)/189,col="blue", ylab="fréquence")
```

```
table(poids$LOW)
barplot(table(poids$LOW)/189,col="blue", ylab="fréquence")



"test khi-2 d'indépendance entre LWO et les autres variables qualitatives"

test1<-chisq.test(table(poids$AGE,poids$LOW))
test1

test2<-chisq.test(table(poids$SMOKE,poids$LOW))
test2

test3<-chisq.test(table(poids$RACE,poids$LOW))
test3

test5<-chisq.test(table(poids$UI,poids$LOW))
test5

"test exact de fisher"

test4<-fisher.test(table(poids$HT,poids$LOW))
test4



"tests de normalité des variables quantitatives à 2 échantillons classés par LOW"
lwt1<-poids$LWT[poids$LOW == "low non"]
ks.test(lwt1,"pnorm",mean(lwt1), sd(lwt1))
lwt2<-poids$LWT[poids$LOW == "low oui"]
ks.test(lwt2,"pnorm",mean(lwt2), sd(lwt2))   "p-value = 0.16"
```

```
fvt1<-poids$FVT[poids$LOW == "low non"]
ks.test(fvt1,"pnorm",mean(fvt1), sd(fvt1))
fvt2<-poids$FVT[poids$LOW == "low oui"]
ks.test(fvt2,"pnorm",mean(fvt2), sd(fvt2))

ptl1<-poids$PTL[poids$LOW == "low non"]
ks.test(ptl1,"pnorm",mean(ptl1), sd(ptl1))
ptl2<-poids$PTL[poids$LOW == "low oui"]
ks.test(ptl2,"pnorm",mean(ptl2), sd(ptl2))

"tests de comparaisons de Wilcoxon"

wilcox.test(poids$LWT~poids$LOW)
wilcox.test(poids$AGE~poids$LOW)
wilcox.test(poids$PTL~poids$LOW)
wilcox.test(poids$FVT~poids$LOW)

"matrice de corrélation"

"codes pour enlever créer une matrice contenant seulement les colonnes des variables quantitatives"
mat_num<-poids
mat_num<-mat_num[,-3]
mat_num<-mat_num[,-3]
mat_num<-mat_num[,-4]
mat_num<-mat_num[,-4]
mat_num<-mat_num[,-6]
mat_num
```

```
"matrice contenant seulement les variables quantitatives"

cor(mat_num,method ="spearman")

"régression logistique binaire, LOW en fonction des autres variables qualitatives"

reg1 <- glm(poids$LOW ~poids$HT,
        data = poids, family = binomial(logit))

reg2<- glm(poids$LOW ~poids$SMOKE,
        data = poids, family = binomial(logit))

reg3<- glm(poids$LOW ~poids$UI,
        data = poids, family = binomial(logit))


"Odds ratio : "

exp(coef(reg1))
exp(coef(reg2))
exp(coef(reg3))

"boxplots de BWT selon les classes des variables qualitatives significatives"

boxplot(poids$BWT~poids$HT,data=poids)
boxplot(poids$BWT~poids$SMOKE,data=poids)
boxplot(poids$BWT~poids$UI,data=poids)

"nuage de points de BWT en fonction de LWT et en fonction de PTL"

plot(x=poids$LWT, y=poids$BWT)
```

```
abline(h=2.5, col="blue", lwd=3)

plot(x=poids$PTL, y=poids$BWT)
abline(h=2.5, col="blue", lwd=3)

"régression logistique de LOW en fonction de PTL et LWT et représentations graphiques des effets"

modelcomplet<-glm(poids$LOW ~ poids$LWT + poids$PTL + poids$LWT:poids$PTL, data=poids,
family=binomial)
model_des<-step(modelcomplet,direction="backward")
model_des

"représentations graphiques de l'effet PTL"

regptl=glm(poids$LOW~poids$PTL, family=binomial(link=logit))
summary(regptl)

logit_ypredit = 0.8018*poids$PTL-0.9642
ypredit=exp(logit_ypredit)/(1+ exp(logit_ypredit))

plot(poids$PTL,poids$LOW)
o=order(poids$PTL)
points(poids$PTL[o],ypredit[o], col="red", type="l", lwd=2)

"représentations graphiques de l'effet LWT"

reglwt=glm(poids$LOW~poids$LWT, family=binomial(link=logit))
summary(reglwt)

logit_ypredit = -0.031*poids$LWT+0.99831
ypredit=exp(logit_ypredit)/(1+ exp(logit_ypredit))
```

```
plot(poids$LWT,poids$LOW)
points(poids$LWT,ypredit, col="red")

plot(poids$LWT,poids$LOW)
o=order(poids$LWT)
points(poids$LWT[o],ypredit[o], col="red", type="l", lwd=2)

"acm"

install.packages("FactoMineR")
library(FactoMineR)

"on enlève les variables non selectionnées pour l'acm"

mat_quali<-poids
mat_quali<-mat_quali[,-1]
mat_quali<-mat_quali[,-1]
mat_quali<-mat_quali[,-1]
mat_quali<-mat_quali[,-2]
mat_quali<-mat_quali[,-4]
mat_quali<-mat_quali[,-4]
mat_quali

res.mca<-MCA(mat_quali, ncp = 2, quali.sup = 4)
summary(res.mca)

plot(res.mca,label=c("var","quali.sup"))
```