Université Grenoble-Alpes
Master Statistique
UE Biostatistique

TD/TP 1 - Régression logistique

EXERCICE 1

On s'intéresse au rôle de différents facteurs dans la survenue de cancer chez l'enfant. Le tableau 1 présente la distribution de cancer de l'enfant selon le statut tabagique chez des mères n'ayant pas bu d'alcool pendant la grossesse.

Consommation de tabac	Non-Malades	Malades
Non-Fumeuses	a=856	b=399
Fumeuses	c=213	d=86

TABLE 1 – Distribution de cancer de l'enfant selon le statut tabagique des mères n'ayant pas bu de l'alcool pendant la grossesse.

1. Identifiez la variable expliquée Y et la variable explicative X_1, ainsi que les valeurs qu'elles peuvent prendre.
2. Exprimez l'OR associé à X_1 en fonction des probabilités conditionnelles de Y sachant X_1, puis en fonction de a,b,c et d.
3. Dans *R*, définissez les valeurs de a, b, c et d, et calculez l'OR associé à X_1. Interpréter.
4. Donnez l'expression analytique du modèle envisagé.
5. Sachant que $odd(X = x) = exp(\beta x)$ et que $var(\beta) = 1/a + 1/b + 1/c + 1/d$, en déduire la valeur de β et de la variance de β.
6. Calculez l'intervalle de confiance de β puis de l'OR. Conclure sur sa significativité.
7. Après avoir posées les hypothèses H_0 et H_1 (en fonction de β et de OR), calculez la statistique du test de Wald pour la variable X_1. Conclure sur sa significativité.

   On obtient par la suite des données de cancer de l'enfant pour des femmes ayant bu pendant la grossesse, présentées dans le tableau 2.

Consommation de tabac	Non-Malades	Malades
Non-Fumeuses	272	88
Fumeuses	76	47

TABLE 2 – Distribution de cancer de l'enfant selon le statut tabagique des mères ayant bu de l'alcool pendant la grossesse.

8. Calculez l'OR associé à X_1 et son IC pour les femmes ayant bu de l'alcool pendant la grossesse. Conclure sur sa significativité.
9. Reproduire le tableau ci dessus que nous aurions obtenu si nous n'avions pas collecté les données pour la variable X_2 "consommation d'alcool" chez les femmes enceintes. Calculez l'OR associé à X_1 et son IC. Conclure sur sa significativité.
10. Conclure sur l'effet de la prise en compte de la variable "consommation d'alcool" dans cette étude.

    On regarde maintenant l'incidence des seuls cancers de types leucémique, pour les enfants de mères ayant bu pendant leur grossesse. Le tableau 3 présente ces résultats.

Consommation de tabac	Non-Malades	Malades
Non-Fumeuses	360	0
Fumeuses	96	27

TABLE 3 – Distribution de leucémie de l'enfant selon le statut tabagique des mères ayant bu de l'alcool pendant la grossesse.

11. Calculez $OR_{1/0}$ et $OR_{0/1}$ associé au risque de leucémie sur cette population. Qu'en concluez vous ?

EXERCICE 2

Les données de l'exercice 1 ont été analysées par modèles logistiques à l'aide de deux modèles, appelés Mod1 et Mod2. Les résultats des modèles sont regroupés dans le tableau 4.

Mod1			Mod2		
-2LogVraisemblance=2501.291			-2LogVraisemblance=2492.344		
Variable	β	$Std(\beta)$	Variable	β	$Std(\beta)$
Intercept	-0.8049	0.0595	Intercept	-0.7633	0.0606
X_1	0.0748	0.1183	X_1	-0.1436	0.1414
X_2	-0.1613	0.1155	X_2	-0.3651	0.1368
			X_1*X_2	0.7915	0.2636

TABLE 4 – Résultats des modèles logistiques.

Le codage des variables est le suivant :
- Y : 0=non-malade, 1=cancer pédiatrique
- X_1 : 0=non-fumeuse, 1=fumeuse
- X_2 : 0=non-buveuse, 1=buveuse.

1. Donnez les expressions analytiques des 2 modèles.
2. Exprimez l'OR associé à X_1 en utilisant l'expression analytique du modèle Mod1.
3. A partie des résultats du modèle Mod1, calculez l'OR (ajusté sur la consommation d'alcool) associé à X_1 et l'OR (ajusté sur la consommation de tabac) associé à X_2.
4. Pour ces OR, calculez leur statistique de test pour le test de significativité de Wald. Conclure sur leur significativité.
5. A partir des résultats du modèle Mod2, calculez l'OR associé à X_1 chez les sujets dont les mères sont buveuses, et celui chez les sujets dont les mères sont non-buveuses. Comparez aux résultats obtenus aux questions 1.3 et 1.8.
6. On veut tester si les ORs associés à X_1 diffèrent selon la consommation d'alcool. Quel paramètre doit-on regarder ? Effectuez le test de significativité de ce paramètre selon la méthode de Wald et selon la méthode du rapport de Vraisemblance. La conclusion est-elle la même qu'à l'exercice 1 ?
7. Exprimez la probabilité de cancer pédiatrique $P(Y = 1|X)$ en utilisant l'expression du modèle Mod2.
8. A l'aide des données du modèle Mod1, calculez la probabilité de cancer pédiatrique pour un sujet dont la mère fumait et ne consommait pas d'alcool pendant la grossesse.

EXERCICE 3 (TP sous R)

1. Chargez la base de données *grossesse.csv* sous R en utilisant la fonction *read.table*. Identifiez les variables par leur nom de colonne.
2. Affichez les statistiques descriptives de la table en utilisant la fonction *summary*. Quel indicateur descriptif manque ?
3. Vérifiez que les effectifs sont bien ceux décrits dans les Tables 1 et 2.
4. Appliquez un modèle de régression associant la survenue de cancer au statut tabagisme sur la population entière, en utilisant la fonction *glm()*. Quelle est la valeur du coefficient associé ? Retrouve t'on l'odds-ratio calculé en 1.9 ?
5. Calculez les intervalles de confiance à l'aide de la fonction *confint.default()*. Comparez avec celui calculé en 1.9.
6. Appliquez les modèles logistiques appropriés et retrouver les résultats du tableau 4.