Samir Chergui

Le 22 Avril 2019

STA110 : Devoir N°2

A l'attention de Mr Jaupi.

Jeu de donnée à caractère personnel utilisé.

Dans le cadre de cet exposé, nous allons étudier une population de 1658 individus plus précisément 1658 conducteurs et leur mortalité compte tenu de leur âge, de leur alcoolisation et leur éventuelle utilisation de drogue.

La variable dépendante Y sera la variable « décès », cette variable est une variable dummy codée 1 ou 0 : 1 signifiant la mort du conducteur et 0 signifiant la survie de celui-ci.

Les variables explicatives seront :

→L'âge au moment de l'accident

→Consommation d'alcool : variable dummy (=1 si non alcoolisé et =2 si alcoolisé)

→Consommation de drogues : variable dummy (=0 si non drogué et =1 si drogué)

L'objectif de cette étude sera de savoir si il existe un lien entre consommation de drogues, d'alcool et âge de l'individu et son « décès » puis le cas échéant tenir compte des liens significatifs pour établir un modèle permettant d'estimer la probabilité d'occurrence du décès ou de la survie compte tenu des caractéristiques du conducteur.

Enfin, nous chercherons à savoir les caractéristiques des conducteurs ayant la valeur 1 pour la variable « Décès ».

Pour essayer d'élaborer une réponse à ce problème, nous utiliserons une régression logistique qui prendra en compte l'ensemble des variables explicatives, nous utiliserons le logiciel Statgraphics.

Après chargement du fichier, nous sélectionnons relier, données de type attribut puis régression logistique. Nous considérons la variable décès comme la variable à expliquer, la variable âge comme un facteur quantitatif et les variables consommation d'alcool et de drogues comme étant des variables qualitatives.

Nous avons les résultats suivants :

Régression logistique - Décès

Variable à expliquer: Décès
Facteurs:
  age_pendant_accident
  consommation_alcool
  utilisation_de_drogues

Nombre d'observations: 1658

Modèle estimé de régression (Maximum de vraisemblance)

		Erreur	*Rapports des chances*
Paramètre	*Estimation*	*type*	*estimées*
CONSTANTE	-7,47532	0,606257	
age_pendant_accident	0,123716	0,0105028	1,13169
consommation_alcool=1	-0,450978	0,133796	0,637005
utilisation_de_drogues=0	-0,0514352	0,124831	0,949865

Analyse de l'écart

Source	*Ecart*	*Ddl*	*Proba.*
Modèle	215,095	3	0,0000
Résidu	1584,33	1654	0,8882
Total (corr.)	1799,43	1657	

Le modèle ici convient car la p-value est égale à 0,0000 ; le résidu (p-value égale à 88,82%) montre qu'il n'y a pas d'écart à l'ajustement.

Le pourcentage d'écart expliqué par le modèle est de 11,9535 %
Pourcentage ajusté = 11,5089

Tests sur les rapports de vraisemblance

Facteur	*Khi-carré*	*Ddl*	*Proba.*
age_pendant_accident	158,203	1	0,0000
consommation_alcool	11,1664	1	0,0008
utilisation_de_drogues	0,169756	1	0,6803

Décès = exp(eta)/(1+exp(eta))

où

eta = -7,47532 + 0,123716*age_pendant_accident - 0,450978*consommation_alcool=1 - 0,0514352*utilisation_de_drogues=0

Le modèle est donc :

Décès = exp(-7,47532 + 0,123716*age_pendant_accident - 0,450978*consommation_alcool=1 - 0,0514352*utilisation_de_drogues=0)/(1+exp(-7,47532 + 0,123716*age_pendant_accident - 0,450978*consommation_alcool=1 - 0,0514352*utilisation_de_drogues=0))

On constate que les p-values de l'âge et de la consommation d'alcool sont hautement significatives alors que la p-value de l'utilisation de drogues (68,03%) montre que ce facteur n'est pas significatif au niveau de confiance de 95% et que par conséquent il peut être envisageable de le retirer.

Essayons d'améliorer le pourcentage ajusté en utilisant la procédure backward (avec des pas de 0,05) : On commence par un modèle complet c'est-à-dire un modèle qui prends en compte l'ensemble des facteurs puis un facteur est retiré si il y a une augmentation du R carré ajusté.

En considérant l'ensemble des étapes, nous obtenons les résultats suivants :

Régression logistique - Décès

Variable à expliquer: Décès
Facteurs:
  age_pendant_accident
  consommation_alcool
  utilisation_de_drogues
Nombre d'observations: 1658

Modèle estimé de régression (Maximum de vraisemblance)

		Erreur	*Rapports des chances*
Paramètre	*Estimation*	*type*	*estimées*

CONSTANTE	-7,49412	0,604524	
age_pendant_accident	0,123598	0,0104952	1,13156
consommation_alcool=1	-0,452255	0,133734	0,636192

Analyse de l'écart

Source	*Ecart*	*Ddl*	*Proba.*
Modèle	214,925	2	0,0000
Résidu	1584,5	1655	0,8909
Total (corr.)	1799,43	1657	

Le modèle ici convient car la p-value est égale à 0,0000 ; le résidu (p-value égale à 89,09%) montre qu'il n'y a pas d'écart à l'ajustement.

Pourcentage d'écart expliqué par le modèle = 11,9441
Pourcentage ajusté = 11,6106

Le pourcentage ajusté s'est amélioré donc ce modèle a une meilleure puissance explicative que le précédent.

Tests sur les rapports de vraisemblance

Facteur	*Khi-carré*	*Ddl*	*Proba.*
age_pendant_accident	158,056	1	0,0000
consommation_alcool	11,2408	1	0,0008

Les deux facteurs sont hautement significatifs car les p-values sont proches de zéro.

Sélection pas à pas des facteurs
Méthode: sélection descendante
P-en-entrée: 0,05
P-en-sortie: 0,05

Etape 0:
3 facteur(s) dans le modèle. 1654 ddl pour l'erreur.
Pourcentage d'écart expliqué = 11,95% Pourcentage ajusté = 11,51%

Etape 1:
Suppression du facteur utilisation_de_drogues avec P-en-sortie = 0,680328
2 facteur(s) dans le modèle. 1655 ddl pour l'erreur.
Pourcentage d'écart expliqué = 11,94% Pourcentage ajusté = 11,61%

Nous avons :

Décès = exp(eta)/(1+exp(eta))

où

eta = -7,49412 + 0,123598*age_pendant_accident - 0,452255*consommation_alcool=1

L'équation du modèle ajusté final devient :

Décès = exp(-7,49412 + 0,123598*age_pendant_accident - 0,452255*consommation_alcool=1)/(1+exp(-7,49412 + 0,123598*age_pendant_accident - 0,452255*consommation_alcool=1))

La probabilité d'occurrence du décès augmente donc lorsque l'individu est relativement peu agé et qu'il n'a pas consommé d'alcool (car la variable 1 signifie non-consommation d'alcool.)

Tableau des prévisions inverses pour age_pendant_accident
consommation_alcool=1
utilisation_de_drogues=0

		LC inf. à 95,0%	*LC sup. à 95,0%*
Pourcentage	*age_pendant_accident*		
0,1	8,41127	-0,687838	14,9432
0,5	21,4653	14,9314	26,1756
1,0	27,1141	21,6783	31,0483
2,0	32,8043	28,4592	35,9719
3,0	36,1678	32,4549	38,8949

4,0	38,5792	35,31	41,0
5,0	40,4693	37,5397	42,6581
6,0	42,03	39,3735	44,0348
7,0	43,3637	40,9334	45,2182
8,0	44,5316	42,2924	46,2615
9,0	45,5729	43,4974	47,1986
10,0	46,5148	44,5803	48,053
15,0	50,2578	48,7879	51,5446
20,0	53,0758	51,7989	54,3303
25,0	55,4034	54,1519	56,765
30,0	57,4367	56,1149	58,9844
35,0	59,2835	57,8404	61,0577
40,0	61,0114	59,4197	63,0329
45,0	62,6684	60,9114	64,9494
50,0	64,2919	62,3577	66,8427
55,0	65,9155	63,793	68,7471
60,0	67,5724	65,2494	70,699
65,0	69,3004	66,7614	72,7414
70,0	71,1472	68,3716	74,9299
75,0	73,1805	70,1392	77,3448
80,0	75,5081	72,1575	80,1142
85,0	78,3261	74,5958	83,4726
90,0	82,0691	77,828	87,9396
91,0	83,0109	78,6404	89,0645
92,0	84,0523	79,5383	90,3087
93,0	85,2201	80,5449	91,7043
94,0	86,5539	81,694	93,2986
95,0	88,1146	83,0382	95,1648
96,0	90,0047	84,6654	97,4255
97,0	92,4161	86,7404	100,311
98,0	95,7796	89,6333	104,336
99,0	101,47	94,5244	111,15
99,5	107,119	99,3774	117,916
99,9	120,173	110,586	133,56

On observe grâce à ce tableau des prévisions inverses que les individus ayant un âge inférieur à 64,2919 ans sont susceptibles d'avoir une valeur de « Décès » égale à zéro et inversement les individus plus âgés seraient susceptibles de se faire attribuer une valeur de 1 pour « Décès » (Ceci en présupposant que les conducteurs n'ont consommés ni drogue ni alcool).