REGRESSION LOGISTIQUE

ASSURES BELGES

Cette étude concerne l'assurance automobile. L'échantillon est constitué de 1106 assurés belges observés en 1992 et répartis en deux groupes.
Les assurés qui n'ont eu aucun accident dans l'année, qui sont au nombre de 556 (modalité « 0 sin » de la variable SINIST)
Les assurés ayant eu au moins un sinistre dans l'année, qui sont au nombre de 550 (modalité « >1 sin » de la variable SINIST)
On entend par assuré une personne physique ou morale.

L'objectif de l'étude est de modéliser la probabilité de SINIST='> 1 sin'.

Les calculs ont été effectuées avec la procédure LOGISTIC de la version française 9.1 du logiciel SAS. Cette version a la particularité d'alterner français et anglais au niveau des sorties.

Les variables explicatives utilisées dans cette étude concernent le souscripteur et le véhicule :

- Le sexe (3 modalités) : SEXE
  Sexe féminin
  Sexe masculin
  Sexe autre (Société)

- Le code d'usage : CUSAG
  Privé
  Professionnel

- L'année de construction du véhicule : DCONS2M
  1933-1989 DCOS
  1990-1991 DCOS

- L'année de naissance de l'assuré AGE3M
  1890-1949
  1950-73
  Naiss ???

- Le degré de bonus-malus de l'année précédente (1991) : BM2M_1
  Autres B-M (-1) pas de bonus en 1991
  B-M 1 (-1) bonus en 1991

- La date d'effet de la police : DPOLI2M
  <86 Police antérieure à 1986
  autres polices

- La langue de rédaction du contrat : CLANG
  Lang franç. Français
  Lang néerland. Néerlandais

- Le code postal : CPOST2M
  Bruxelles
  Autres

- La puissance du véhicule : PUIS2M
  10-39 Puis
  40-349 Puis

1) Dans cette première phase de l'étude on cherche à modéliser **la probabilité de SINIST='> 1 sin'** en fonction de neuf variables explicatives :
   CUSAG SEXE CLANG AGE3M CPOST2M BM2M_1 PUIS2M DPOLI2M DCONS2M

   On réalise dans ce but une régression logistique (lien logit) utilisant ces neuf variables.

   Commentez les résultats présentés dans le tableau 1 (modèle complet). Que proposez-vous pour la suite de l'étude ?

   D'autres approches que celle proposée seraient-elles envisageables ?

2) Une sélection de variables (option STEPWISE de la procédure LOGISTIC) a été mise en œuvre (TABLEAU 2). Quel modèle proposez vous de retenir ? Il est demandé d' utiliser plusieurs critères pour justifier ce choix. Précisez comment s'interprète le test du khi-deux résiduel disponible à chaque étape de la sélection.

3) Une régression logistique sur sept variables est proposée dans le tableau 3. Ecrire le modèle estimé. Interpréter les coefficients. Comment utilise-ton les rapports de côte (ODDS RATIO) ?

4) Une régression logistique sur sept variables, utilisant l'option param = glm, qui permet un codage des effets comme dans la procédure GLM (c'est à dire que la dernière catégorie est prise en référence et mise à zéro) est proposée dans le tableau 4 . Ecrire le modèle estimé.
Calculer la probabilité de sinistre pour un assuré ayant les caractéristiques suivantes :

CUSAG = Privé
AGE3M = 1890-1949
CPOST2M = Bruxelles
BM2M_1 = Autres B-M (-1)
PUIS2M = 10-39
DPOLI2M = <86 Police
DCONS2M = 33-89 DCOS

Vérifiez ce résultat en utilisant le modèle défini dans le tableau 3.

5) Dans le tableau 4, qu'apportent les statistiques fournies sous la rubrique : « Association des probabilités prédites et des réponses observées » ?
Indiquer l'utilisation de la table de classification présentant la sensibilité, la spécificité…Quel graphique pourrait synthétiser ce type d'informations ?

ANNEXE 1 : ASSURES BELGES

TABLEAU 1 :MODELE COMPLET

The LOGISTIC Procedure

Informations sur le modèle

Data Set	IN.ASSUR
Response Variable	SINIST
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1106
Number of Observations Used	1106

Profil de réponse

Valeur ordonnée	SINIST	Fréquence totale
1	> 1 sin	550
2	0 sin	556

État de convergence du modèle

Convergence criterion (GCONV=1E-8) satisfied.

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	848.799
SC	1540.218	908.901
-2 Log L	1533.209	824.799

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	708.4096	11	<.0001
Score	592.8368	11	<.0001
Wald	370.4675	11	<.0001

TABLEAU 1 :MODELE COMPLET (SUITE)

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
CUSAG	1	6.9000	0.0086
SEXE	2	1.4089	0.4944
CLANG	1	0.3124	0.5762
AGE3M	2	45.2598	<.0001
CPOST2M	1	18.1904	<.0001
BM2M_1	1	159.2260	<.0001
PUIS2M	1	8.3037	0.0040
DPOLI2M	1	5.9536	0.0147
DCONS2M	1	40.6677	<.0001

Association des probabilités prédites et des réponses observées

Percent Concordant	90.5	Somers' D	0.816
Percent Discordant	8.9	Gamma	0.822
Percent Tied	0.7	Tau-a	0.408
Pairs	305800	c	0.908

TABLEAU 2 : SELECTION DE VARIABLES

Stepwise Selection Procedure

Step 0. Intercept entered:

État de convergence du modèle
Convergence criterion (GCONV=1E-8) satisfied.
-2 Log L = 1533.209

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
592.8368	11	<.0001

Step 1. Effect BM2M_1 entered:

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	996.127
SC	1540.218	1006.144
-2 Log L	1533.209	992.127

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	541.0821	1	<.0001
Score	495.2161	1	<.0001
Wald	399.7544	1	<.0001

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
177.1136	10	<.0001

NOTE: No effects for the model in Step 1 are removed.

Step 2. Effect DCONS2M entered:

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	934.893
SC	1540.218	949.918
-2 Log L	1533.209	928.893

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	604.3162	2	<.0001
Score	529.6882	2	<.0001
Wald	374.8460	2	<.0001

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
105.7275	9	<.0001

NOTE: No effects for the model in Step 2 are removed.

Step 3. Effect AGE3M entered:

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	889.217
SC	1540.218	914.260
-2 Log L	1533.209	879.217

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	653.9916	4	<.0001
Score	560.3511	4	<.0001
Wald	374.7194	4	<.0001

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
53.6558	7	<.0001

NOTE: No effects for the model in Step 3 are removed.

Step 4. Effect CPOST2M entered:

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	870.102
SC	1540.218	900.153
-2 Log L	1533.209	858.102

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	675.1069	5	<.0001
Score	572.4280	5	<.0001
Wald	371.7980	5	<.0001

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
32.9625	6	<.0001

NOTE: No effects for the model in Step 4 are removed.

Step 5. Effect CUSAG entered:

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	855.564
SC	1540.218	890.624
-2 Log L	1533.209	841.564

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	691.6450	6	<.0001
Score	582.6443	6	<.0001
Wald	370.8872	6	<.0001

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
16.5984	5	0.0053

NOTE: No effects for the model in Step 5 are removed.

Step 6. Effect PUIS2M entered:

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	848.583
SC	1540.218	888.651
-2 Log L	1533.209	832.583

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	700.6259	7	<.0001
Score	587.4784	7	<.0001
Wald	369.4314	7	<.0001

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
7.8893	4	0.0957

NOTE: No effects for the model in Step 6 are removed.

Step 7. Effect DPOLI2M entered:

Statistiques d'ajustement du modèle

	Coordonnée à l'origine

Critère	Coordonnée à l'origine uniquement	et covariables
AIC	1535.209	844.462
SC	1540.218	889.539
-2 Log L	1533.209	826.462

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	706.7468	8	<.0001
Score	591.3107	8	<.0001
Wald	369.8239	8	<.0001

Test du Khi 2 résiduel

Khi 2	DF	Pr > Khi 2
1.6345	3	0.6516

NOTE: No effects for the model in Step 7 are removed.

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Récapitulatif sur la sélection séquentielle

Étape	Effet Saisi	Effet Supprimé	DF	Nombre dans	Khi 2 du score	Khi 2 de Wald	Pr > Khi 2
1	BM2M_1		1	1	495.2161		<.0001
2	DCONS2M		1	2	62.0832		<.0001
3	AGE3M		2	3	50.3087		<.0001
4	CPOST2M		1	4	21.4766		<.0001
5	CUSAG		1	5	16.2115		<.0001
6	PUIS2M		1	6	8.8740		0.0029
7	DPOLI2M		1	7	6.2583		0.0124

Récapitulatif sur la sélection séquentielle

Étape	Libellé de variable
1	Bonus-malus Année -1 (2 mod) - GBM1
2	Année de construction du véhicule (2 mod) - DCOS 38-39
3	Age de l assuré (3 mod) - DNAI 8-9
4	Code postal souscripteur (2 mod) - POSS2 17-18
5	Code usage - CUSA 5-6
6	Puissance du véhicule (2 mod) - PUIS 32-33
7	Date effet Police (2 mod) - DPEP 26-27

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
CUSAG	1	11.4673	0.0007
AGE3M	2	44.0929	<.0001
CPOST2M	1	22.0972	<.0001
BM2M_1	1	165.0321	<.0001
PUIS2M	1	8.7026	0.0032
DPOLI2M	1	6.2087	0.0127
DCONS2M	1	41.7697	<.0001

TABLEAU 3 : MODELE à 7 variables

The LOGISTIC Procedure

Informations sur le modèle

Data Set	IN.ASSUR	
Response Variable	SINIST	Sinistralité RC - SNB11 2-3
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	1106
Number of Observations Used	1106

Profil de réponse

Valeur ordonnée	SINIST	Fréquence totale
1	> 1 sin	550
2	0 sin	556

Probability modeled is SINIST='> 1 sin'.

Informations sur le niveau de classe

Classe	Valeur	Variables de création	
CUSAG	Privé	1	
	Profess.	-1	
AGE3M	1890-1949	1	0
	1950-73	0	1
	Naiss ???	-1	-1
CPOST2M	Bruxelles	1	
	Autres	-1	
BM2M_1	Autres B-M (-1)	1	
	B-M 1 (-1)	-1	
PUIS2M	10-39 Puis	1	
	40-349 Puis	-1	
DPOLI2M	<86 Police	1	
	autres polices	-1	
DCONS2M	33-89 DCOS	1	
	90-91 DCOS	-1	

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	844.462
SC	1540.218	889.539
-2 Log L	1533.209	826.462

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	706.7468	8	<.0001
Score	591.3107	8	<.0001
Wald	369.8239	8	<.0001

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
CUSAG	1	11.4673	0.0007
AGE3M	2	44.0929	<.0001
CPOST2M	1	22.0972	<.0001
BM2M_1	1	165.0321	<.0001
PUIS2M	1	8.7026	0.0032
DPOLI2M	1	6.2087	0.0127
DCONS2M	1	41.7697	<.0001

Analyse des estimations de la vraisemblance maximum

Paramètre		DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept		1	0.6136	0.1663	13.6201	0.0002
CUSAG	Privé	1	-0.4183	0.1235	11.4673	0.0007
AGE3M	1890-1949	1	-0.3112	0.1320	5.5600	0.0184
AGE3M	1950-73	1	0.9434	0.1443	42.7753	<.0001
CPOST2M	Bruxelles	1	0.4505	0.0958	22.0972	<.0001
BM2M_1	Autres B-M (-1)	1	1.2293	0.0957	165.0321	<.0001
PUIS2M	10-39 Puis	1	-0.3673	0.1245	8.7026	0.0032
DPOLI2M	<86 Police	1	-0.2510	0.1007	6.2087	0.0127
DCONS2M	33-89 DCOS	1	-0.6738	0.1043	41.7697	<.0001

Estimations des rapports de cotes

Effet		Point Estimate	95% Limites de confiance de Wald	
CUSAG	Privé vs Profess.	0.433	0.267	0.703
AGE3M	1890-1949 vs Naiss ???	1.379	0.883	2.151
AGE3M	1950-73 vs Naiss ???	4.834	2.971	7.866
CPOST2M	Bruxelles vs Autres	2.462	1.691	3.585
BM2M_1	Autres B-M (-1) vs B-M 1 (-1)	11.689	8.033	17.010
PUIS2M	10-39 Puis vs 40-349 Puis	0.480	0.294	0.781
DPOLI2M	<86 Police vs autres polices	0.605	0.408	0.898
DCONS2M	33-89 DCOS vs 90-91 DCOS	0.260	0.173	0.391

Association des probabilités prédites et des réponses observées

Percent Concordant	90.3	Somers' D	0.817
Percent Discordant	8.6	Gamma	0.827
Percent Tied	1.2	Tau-a	0.409
Pairs	305800	c	0.908

TABLEAU 4 : MODELE à 7 variables Procédure logistic option param = GLM

Informations sur le modèle

Data Set	IN.ASSUR	
Response Variable	SINIST	Sinistralité RC - SNB11 2-3
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	1106
Number of Observations Used	1106

Profil de réponse

Valeur ordonnée	SINIST	Fréquence totale
1	> 1 sin	550
2	0 sin	556

Probability modeled is SINIST='> 1 sin'.

Informations sur le niveau de classe

Classe	Valeur	Variables de création		
CUSAG	Privé	1	0	
	Profess.	0	1	
AGE3M	1890-1949	1	0	0
	1950-73	0	1	0
	Naiss ???	0	0	1
CPOST2M	Autres codes	1	0	
	Bruxelles	0	1	
BM2M_1	Autres B-M (-1)	1	0	
	B-M 1 (-1)	0	1	
PUIS2M	10-39 Puis	1	0	
	40-349 Puis	0	1	
DPOLI2M	<86 Police	1	0	
	autres polices	0	1	
DCONS2M	33-89 DCOS	1	0	
	90-91 DCOS	0	1	

The LOGISTIC Procedure

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	1535.209	844.462
SC	1540.218	889.539
-2 Log L	1533.209	826.462

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	706.7468	8	<.0001
Score	591.3107	8	<.0001
Wald	369.8239	8	<.0001

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
CUSAG	1	11.4673	0.0007
AGE3M	2	44.0929	<.0001
CPOST2M	1	22.0972	<.0001
BM2M_1	1	165.0321	<.0001
PUIS2M	1	8.7026	0.0032
DPOLI2M	1	6.2087	0.0127
DCONS2M	1	41.7697	<.0001

Analyse des estimations de la vraisemblance maximum

Paramètre		DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept		1	0.9130	0.3623	6.3515	0.0117
CUSAG	Privé	1	-0.8367	0.2471	11.4673	0.0007
CUSAG	Profess.	0	0	.	.	.
AGE3M	1890-1949	1	0.3211	0.2270	2.0001	0.1573
AGE3M	1950-73	1	1.5757	0.2484	40.2329	<.0001
AGE3M	Naiss ???	0	0	.	.	.
CPOST2M	Autres codes	1	-0.9010	0.1917	22.0972	<.0001
CPOST2M	Bruxelles	0	0	.	.	.
BM2M_1	Autres B-M (-1)	1	2.4587	0.1914	165.0321	<.0001
BM2M_1	B-M 1 (-1)	0	0	.	.	.
PUIS2M	10-39 Puis	1	-0.7347	0.2490	8.7026	0.0032
PUIS2M	40-349 Puis	0	0	.	.	.
DPOLI2M	<86 Police	1	-0.5021	0.2015	6.2087	0.0127
DPOLI2M	autres polices	0	0	.	.	.
DCONS2M	33-89 DCOS	1	-1.3476	0.2085	41.7697	<.0001
DCONS2M	90-91 DCOS	0	0	.	.	.

Estimations des rapports de cotes

Effet		Point Estimate	95% Limites de confiance de Wald	
CUSAG	Privé vs Profess.	0.433	0.267	0.703
AGE3M	1890-1949 vs Naiss ???	1.379	0.883	2.151
AGE3M	1950-73 vs Naiss ???	4.834	2.971	7.866
CPOST2M	Autres codes vs Bruxelles	0.406	0.279	0.591
BM2M_1	Autres B-M (-1) vs B-M 1 (-1)	11.689	8.033	17.010
PUIS2M	10-39 Puis vs 40-349 Puis	0.480	0.294	0.781
DPOLI2M	<86 Police vs autres polices	0.605	0.408	0.898
DCONS2M	33-89 DCOS vs 90-91 DCOS	0.260	0.173	0.391

Association des probabilités prédites et des réponses observées

Percent Concordant	90.3	Somers' D	0.817
Percent Discordant	8.6	Gamma	0.827
Percent Tied	1.2	Tau-a	0.409
Pairs	305800	c	0.908

Table de classification

Niveau de prob.	Correct	Pourcentages Sensi-bilité	Spéci-ficité	POS fausse	NEG fausse
0.200	80.7	94.2	67.4	25.9	7.9
0.225	82.0	92.9	71.2	23.8	9.0
0.250	81.9	92.5	71.4	23.8	9.4
0.275	82.7	91.3	74.3	22.2	10.4
0.300	84.4	91.3	77.5	19.9	10.0
0.325	84.4	91.1	77.9	19.7	10.2
0.350	84.2	90.5	77.9	19.8	10.7
0.375	84.4	89.6	79.3	18.9	11.4
0.400	85.3	89.3	81.3	17.5	11.5
0.425	85.4	89.3	81.7	17.2	11.5
0.450	84.9	88.2	81.7	17.4	12.5
0.475	86.0	88.2	83.8	15.7	12.2
0.500	86.2	87.6	84.7	15.0	12.6
0.525	86.1	87.1	85.1	14.8	13.1
0.550	86.3	87.1	85.4	14.5	13.0
0.575	85.6	85.8	85.4	14.6	14.1
0.600	86.1	85.6	86.5	13.7	14.1
0.625	86.3	85.6	86.9	13.4	14.1
0.650	83.8	80.7	86.9	14.1	18.0
0.675	83.5	76.9	89.9	11.7	20.3
0.700	83.5	76.4	90.5	11.2	20.5
0.725	83.2	75.8	90.5	11.3	20.9
0.750	82.9	74.5	91.2	10.7	21.6
0.775	80.5	68.2	92.6	9.9	25.4
0.800	79.6	66.2	92.8	9.9	26.5