Projet : Modélisation de sinistres et de montant de police d'assurance sur données réelles.
Cours L3 MIASHS Projet : Statistiques et simulations S6

A rendre au plus tard le 12 mai 2023, sur cours en ligne sous la forme d'un pdf unique d'une vingtaine de pages (avec graphiques) + annexes éventuelles (graphiques, programmes et résultats d'estimation)

Une compagnie d'assurance a collecté des données sur un portefeuille d'assurés, des données personnelles du ménage (composition du ménage, nombre d'enfants, salaire, région d'habitation, taille d'agglomération etc...), les montants des dépenses du ménage pour se couvrir contre certains risques (Police1 (obligatoire), Police2, Police3) et les montants de sinistres pour trois types de risques. Ces données sont stockées dans le fichier R, assurance.R qu'il vous suffit de sourcer (source("nom fichier")) pour récupérer les données dat (que vous pouvez visualisez sous Rcmd ou avec fix(dat)). Le dictionnaire des variables de la base est donnée en annexe 1. Pour des questions de confidentialité, pour les variables sinistres et polices, l'unité de mesure n'est pas connue (on a appliqué une transformation linéaire arbitraire connue seulement de la compagnie), mais on pourra considérer par exemple qu'il s'agit de dizaines d'euros pour les sinistres et pour les polices. En fait, la compagnie dispose aussi de quelques individus (peu nombreux) n'ayant eu aucun sinistre de type 1. Elle n'a conservé dans un premier temps que les individus ayant eu des sinistres de type 1 strictement positifs et a éliminé les quelques individus n'ayant pas eu aucun sinitre de type 1 . Nous verrons plus tard sur les sinistres de type 2 et 3 (partie 4), les problèmes qui se posent lorsqu'on observe des données avec beaucoup de 0.

Dans la suite, on va chercher à étudier les variables pouvant influencer le montant de sinistres selon les caractéristiques du ménages. En pratique, ce modèle dit de "tarification" est ensuite utilisé pour calculer le montant de la prime pure qu'un individu doit payer pour s'assurer contre un type de sinistres. L'idée sous-jacente étant, qu'à caractéristiques égales, le montant potentiel futur de ses sinistres sera prédit par le modèle et qu'il doit donc payer au minimum le montant prédit de son futur sinistre réparti sur une période.

Partie 1 : Description des données/Tests d'hypothèse

1) Donner quelques statistiques descriptives de quelques variables choisies (moyennes , variances, quartiles, comptage pour les variables qualitatives, box-plot).

2) Tracer sur un même graphique les histogrammes (avec l'option freq=FALSE pour avoir des proportions en ordonnées) et les densités estimées des Sinistres 1 (Sin1) et 2 (Sin2). Effectuer un test de normalité de ces variables. On pourra utiliser un test de Kolmogorov-Smirnov (en R on utilisera la librairie DescTools et le test de KS avec paramêtres estimés lillietest, voir l'aide) ou d'autres tests de normalité (Anderson-Darling). Tracer aussi la loi normale correspondante sur le même graphique (en rouge).

Partie 2 : Modèle linéaire simple.

Un étudiant décide d'expliquer le montant des sinistres de type 1 (Sin1) des ménages en fonction de la variable log(RUC) (log du revenu par unité de consommation du ménage).

1) Etudier ce modèle et dire ce que vous en pensez. La variable log(RUC) est telle significative? Comment interpréter vous le résultat?

2) Tracer sur un même dessin les données et la droite de régression.

3) Estimer le modèle en rajoutant dans le modèle les catégories socio-professionnelles. La variable log(RUC) est-elle encore significative? Pourquoi?

4) Quel modèle choisiriez vous entre un modèle (avec constante) avec seulement log(RUC) comme variable explicative, un modèle avec log(RUC) et pcs, un modèle avec pcs seulement?

Partie 3 : Modèle linéaire multiple

Exclure les 10 derniers individus de la base de données dat et les mettre dans une base à part dat0. En utilisant toutes les ressources du cours, calibrer un modèle faisant intervenir toutes les variables de dat pertinentes pour modéliser les

1) Sinistres de type 1 (Sinistre1) et la Police1 (qui peut dependre de la variable Sinistre1)

On précisera les tests utilisés et comment sont retenues les variables.

3) Pour le/les modèles retenus, tester graphiquement, la normalité des résidus (est-ce utile ici?), l'existence de points abérrants etc...

4) Pour Police1, on liminera les points qui paraissent abérrants et on estimera le modle sans ces individus.

4) En déduire les prédictions des montants des sinistres Sinistre1 pour les individus de la base dat0, et en dduire le montant de Police1 . On donnera ces valeurs dans un tableau avec le montant de leur sinistre rel et de police observé.

Partie 4: Modélisation des zéros dans une régression.

Les variable Sinistre2 et Sinistre3 comportent de nombreuses valeurs avec des zéros, c'est à dire des individus n'ayant pas eu de sinistres de type 2 ou 3. Cette partie a pour but de montrer que inclure les zéros dans la procédure des m.c.o. ou les éliminer peut créer de nombreux biais et de proposer une méthode d'estimation adéquate.

1) On va d'abord montrer par simulations quels sont les problèmes rencontrés avec ce type de données lors de la phase d'estimation. On considère des variables explicatives i.i.d, Z_i de loi $exp(N(0,1))$, $i = 1, ..., n$ (i.e. de loi log-normale $LN(0,1)$) et des variables i.i.d X_i de loi $LN(2,1)$, $i = 1, ..., n$. On va chercher à simuler et à estimer le modèle suivant en deux étapes :

i) La variable dite de sélection est la variable à valeur dans $\{0,1\}$ définie par

$$\delta_i = \left\{ \begin{array}{l} 0 \text{ si } U_i = a + bZ_i + \varepsilon_i < 0 \\ 1 \text{ si } U_i = a + bZ_i + \varepsilon_i > 0 \end{array} \right. \quad \text{où } \varepsilon_i \sim N(0,1)$$

où a et b sont deux paramètres dans $\mathbb{R}$, pour $i = 1..., n$.

δ_i est une variable qui modélise le fait d'avoir un sinistre de type 3. La variable U_i est en fait une variable inobservée qui modélise la propension de l'individu à avoir un sinistre (dépendante de la variable explicative Z_i) : comme on ne connait pas l'échelle de cette utilité, on peut toujours se ramener à supposer que la variance des résidus est 1. Pour fixer les idées, les sinistres 3 sont des accidents automobiles, Z_i est la vitesse moyenne (dans une certaine unité) à laquelle conduit l'individu i. Dans un modèle plus général on pourra rajouter, le sexe du conducteur, son age, son lieu d'habitation (rural/urbain) etc...

ii) la variable $Y_i = \left\{ \begin{array}{c} 0 \quad si \quad \delta_i = 0 \\ \alpha + X_i\beta + \eta_i, \text{ sinon} \end{array} \right\}$ avec η_i i.i.d $N(0, \sigma^2)$. Y_i répresente le montant du sinistre automobile , lorsque l'individu a bien eu un sinistre et est fonction d'une autre variable (par exemple on pourra considérer que X_i est le prix à l'argus du véhicule de l'individu i) .

Montrer que δ_i suit une loi $Ber(1, p_i)$ avec probabilité $p_i = \Phi(a + bZ_i)$.

2) Prendre $a = -0.5$, $b = 1$, $\alpha = 2 * votre\ jour\ de\ naissance$ et $\beta = \frac{2}{votre\ mois\ de\ naissance}$. On prend $\varepsilon_i = \eta_i$ (dans ce cas la dépendence entre les deux étapes est grande). Simuler un jeu de données de taille $n = 100$ de réalisations $(\delta_i, Y_i,\ X_i, Zi)$. Représenter sur un même graphique :

- les données
- la droite exacte $y = \alpha + x\beta$

- la droite de regression des Y_i sur X_i en prenant toutes les données (modèle complet)
- la droite de regression lorsque qu'on ne garde que les Y_i positifs (regressés sur les X_i correspondant) (dit modèle positif).

Commenter ces résultats. Refaire la même chose avec $n = 1000$. Les estimateurs des mco (sur tous l'échantillons et uniquement sur variables positives) vous semblent-ils convergents? Quels sont les problèmes?

3) Simuler 999 échantillons de taille 100 et évaluer le biais et la variance des estimateurs des mco.

4) D'après la question 1), le modèle δ_i est un modèle probit qui peut s'estimer simplement sous R avec la fonction glm(y~x, family=binomial(link="probit"),...). Estimer le sur les données produites en 2) . Obtient-on des estimateurs convergents de a et b?

5) Montrer que si une variable aléatoire Z est $N(0,1)$ alors $E(Z|Z > c) = \frac{\phi(-c)}{\Phi(-c)}$. En déduire la valeur de $E(Z|Z > c)$ lorsque Z est de loi $N(m, \sigma^2)$. On admettra (ou si on est courageux on montrera) que dans un modèle gaussien où

$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix}$ est de loi gaussienne $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right)$, on a Y_i qui vérifie

$$E(Y_i|\delta_i = 1) = \alpha + \beta X_i + \rho\frac{\phi(aZ_i + b)}{\Phi(aZ_i + b)}. \quad (1)$$

Proposer une procédure simple en deux étapes qui permette d'estimer par les mc.o. les paramètres (α, β) et ρ (au moins asymptotiquement) (Aide : on pensera à d'abord estimer a et b par des estimateurs convergents $\widehat{a}_n$ et $\widehat{b}_n$ qu on utilisera dans l'équation (1)). La mettre en oeuvre sur sur une simulation de taille $n = 100$ et $n = 1000$ et comparer aux autres estimations.

6) Le modèle qui a été proposé est en fait un tobit généralisé ou Tobit2. L'estimation en deux étapes décrites précédemment a valu, parmi d'autres travaux, un prix Nobel d'économie à James Heckman. Elle est appelé procédure d'Heckman. Grâce à ces idées, calibrer un modèle pour les sinistres de type 3 en testant la significativité des variables introduites en proposant des variables pour décrire le fait d'avoir un sinistre (on créera la variable delta=1 si Sinistre3>0 et 0 sinon) et les variables explicatives des montants positifs. Interpr̆eter le modèle.

Annexe 1

Dictionnaire de variables
pcs = categories socio-professionnelles
1 = Agriculteurs exploitants
2 = Artisans, commercants, chefs d'entreprises
3 = Cadres et professions intellectuelles superieures
4 = Professions intermediaires
5 = Employes
6 = Ouvriers
7 = Retraites
8 = Autres personnes sans activite professionnelle

cs = categories sociales de revenu
1 = Aisee
2 = Moyenne superieure
3 = Moyenne inferieure
4 = Modeste

reves = Revenu du foyer (estimé dans une tranche)
2000 = 0-4000
4500 = 4000-5000
5500 = 5000-6000
6500 = 6000-7000
7500 = 7000-8000
8500 = 8000-9000
9500 = 9000-10000
11250 = 10000-12500
13750 = 12500-15000
16250 = 15000-17500
18750 = 17500-20000
22500 = 20000-25000
27500 = 25000-30000
50000 = 30000-70000
RUC = Revenu par unités de consommation (variable continue)
(c'est le revenu du ménage divisé non pas par le nombre de personnes dans le foyer mais par une pondération du nombre de personnes, en gros un adulte compte pour 1, un enfant en très bas age 0.2, 0.5 pour des enfant -13 ans etc... selon une échelle qui n'est pas donnée ici).

Regions codée de 1 à 9, 1=Ile de France

Ahabi = Habitat
Communes rurales
Unites urbaines de 2 000 a 9 999 habitants
Unites urbaines de 10 000 a 99 999 habitants

Unites urbaines de 100 000 habitants et plus
Paris + Agglomeration

habi : un codage de la variable Ahabi

Atyph = Statut d'occupation de l'habitation
1 = Proprietaire
2 = Locataire
3 = Non declare

Acompm = Composition du menage
1 = Couple avec enfant(s)
2 = Couple sans enfant
3 = Personne seule
4 = Autre menage

nbpers = Nombre de personnes dans le menage
1
2
3
4
etc...

Nbadulte : nombre d'adultes dans le foyer

Bauto = Possession d'un vehicule automobile
Pas de vehicule
Au moins un vehicule

NSin = nombre de sinistres du ménage dans le passé (depuis l'entrée de l'individu dans la base).

Sin1 : Montant dommage en dizaines d'euros pour les sinistres de type 1
Sin2 : Montant dommage pour les sinistres de type 2
Sin3 : Montant dommage pour les sinistres de type 3

Police 1 : cotisation au titre du complément de police de type 1
Police 2 : cotisation au titre du complément de police de type 2
Police 3 : cotisation au titre du complément de police de type 3