École d'actuariat
Université Laval

ACT-2003 Modèles linéaires en actuariat
Travail pratique I

Chargé de cours : Olivier Côté
Automne 2022

1 Consignes générales

- Date limite pour la remise : **Mercredi le 9 novembre 2022 à 23h59**.
- Le travail doit être effectué en équipe de 4 personnes.
- Date limite pour la formation des équipes sur le site de cours : **Mercredi le 21 septembre 2022 à 23h59**. L'enseignant formera des équipes de quatre avec les étudiant·e·s d'équipes encore incomplètes à cette date.
- Les logiciels R ou Excel (ou les deux) peuvent être utilisés.
- Le rapport en format pdf doit être remis dans la boite de dépôt du site de cours.
- Le rapport, incluant les formules mathématiques, doit être rédigé avec un logiciel de traitement de texte (Word, LaTeX, Rmarkdown, etc.) Un gabarit de rapport produit avec LaTeX est disponible sur le site du cours pour les intéressés.
- Le rapport doit être rédigé de façon structurée. Les graphiques et tableaux doivent tous être expliqués dans le corps du texte.
- 10 % des points sont accordés pour la qualité du français, à raison de 1/2 point par faute d'orthographe ou de grammaire.
- 10 % des points sont accordés pour la qualité de la présentation, et pour la présence d'une page titre appropriée, d'une table des matières et d'une pagination correcte. Les tableaux et graphiques doivent tous avoir un titre. Les textes des graphiques doivent être assez gros pour être lisibles, et les titres et légendes doivent être clairs.

2 Description du projet

L'établissement des provisions liées à des accidents automobiles est essentiel pour qu'une compagnie d'assurance soit solvable. Cela représente un défi important pour la compagnie d'assurance Assurancetourix, pour laquelle vous travaillez.

L'actuaire désigné de votre compagnie a assisté à un congrès où il a entendu parler des grandes promesses de l'analyse prédictive. Il vous demande d'utiliser les méthodes de régression linéaire multiple pour modéliser le montant total payé pour une réclamation en fonction des variables d'indemnisation disponibles au moment de l'ouverture d'un dossier de sinistre.

Vous disposez des informations suivantes, que vous pouvez importer en R directement avec la fonction `read.csv2` :

Tableau 1 – Variables contenues dans le jeu de données.

Variable	Description
`index`	Numéro de police du·de la réclamant·e
`age_permi`	Âge du permi de conduire du réclamant, en mois
`age_veh`	Âge du véhicule accidenté, en années
`genre`	Genre du réclamant
`stat_matri`	Statut matriomonial du réclamant
`categ_socio`	Catégorie socioprofessionnelle (de "`CSP1`" à "`CSP99`")
`usage_veh`	Utilisation habituelle du véhicule accidenté par le réclamant
`age_conduct`	Âge du réclamant, en années
`type_veh`	Type du véhicule accidenté ("`sedan`", "`van`", ...)
`bonus_malus`	Score représentant l'historique de sinistre du réclamant, les valeurs sont numériques et dans l'intervalle $[50, 350[$
`mont_recl`	Montant final de réclamation en dollars
`garage`	Type de garage habituel du réclamant
`source_energ`	Source d'énergie qui permet de faire fonctionner le véhicule accidenté
`valeur_veh`	Valeur du véhicule accidenté en dollars. On ne spécifie pas s'il s'agit de la valeur à l'achat ou actuelle

Au moment où un accident survient, la compagnie doit établir une provision pour couvrir le montant lié à la réclamation. Pour l'instant, le modèle utilisé est simpliste ; la réserve est toujours établie à 2000 $. Votre mandat est d'élaborer un modèle qui aura une meilleure puissance prédictive.

La base de données `MontantReclFr.csv` contient 5323 observations que vous devez utiliser pour ajuster votre modèle. La base de données `MontantReclFr_valid.csv` contient 100 observations, que vous utiliserez pour valider votre modèle dans les sections appropriées.

Votre superviseure vous demande de fournir les prévisions ponctuelles de montant final de réclamation `mont_recl` pour les réclamations du mois passé. Les détails de ces réclamations sont contenus dans le fichier `MontantReclFr_test.csv`. La qualité de votre modèle proposé sera mesurée avec l'erreur quadratique moyenne de prévision sur ces données de test.

Vous devez également fournir un rapport à l'attention de votre superviseure, et y inclure les faits saillants qui seront transmis à l'actuaire désigné.

3 Contenu du rapport

Votre superviseure connaît les rudiments de la régression linéaire multiple. Il n'est donc pas nécessaire d'expliquer la théorie dans votre rapport. Par contre, elle s'attend à lire une analyse complète des données et des résultats, qui comprendra les éléments suivants :

Une page titre : Indiquez votre numéro d'équipe ainsi que toutes les informations pertinentes.

Des faits saillants : Seule cette partie du rapport sera lue par la haute direction de l'actuariat corporatif. Elle doit donc être brève (maximum 250 mots) tout en contenant les informations importantes sur votre analyse et vos conclusions.

Vous pouvez consulter ce lien : Conseils sur l'écriture d'un sommaire exécutif.

Une table des matières

Une analyse univariée des données : Dans cette section, votre superviseure s'attend à lire un résumé clair et concis des données. Cela peut être fait à l'aide de tableaux de fréquence, d'histogrammes ou d'autres types de diagrammes. Les moyennes, écart-types, médiane, minimum et maximum sont des statistiques intéressantes pour des variables continues. Expliquez brièvement les données. Faites les graphiques en nuage de points ou les diagrammes en boîtes à moustaches de la variable endogène en fonction de chacune des variables exogènes à considérer.

Les erreurs doivent déjà être corrigées dans les graphiques.

Note : Si vous identifiez des erreurs dans les données, elles doivent d'abord être corrigées. Mentionnez ces erreurs et la façon dont elles ont été corrigées dans l'annexe A de votre rapport.

Explication du modèle proposé : Dans cette section, présentez les éléments suivants

1. l'équation du modèle proposé ;
2. le traitement des variables qualitatives dans le modèle ;
3. un tableau contenant les estimations des paramètres du modèle et les intervalles de confiance de niveau 95% ;
4. l'interprétation des paramètres du modèle ;
5. l'explication des interactions entre les variables du modèle (s'il y a lieu) ;
6. le coefficient de détermination et son interprétation ;
7. le tableau ANOVA ainsi que des commentaires sur la validité globale de la régression.

Attention : votre superviseure ne veut pas voir toutes les étapes que vous avez dû faire pour atteindre ce modèle que vous préférez. Vous devez donc sélectionner les variables qui sont significatives, ou effectuer les transformations sur les variables (si nécessaire) avant de proposer votre modèle final.

Analyse des résidus : Dans cette section, expliquez si les postulats du modèle de régression linéaire sont vérifiés. Présentez les graphiques et les analyses reliées à la vérification des postulats. Vérifiez s'il y a un problème de multicolinéarité.

Prévisions : Dans votre rapport, présentez les prévisions (en dollars) et les intervalles de confiance sur la prévision pour les cinq premières lignes du fichier `MontantReclFr_valid.csv`. Expliquez comment vous les avez obtenus, formules à l'appui. Utilisez un niveau de confiance de 95 %. Présentez le graphique en nuage de points des prévisions en fonction des observations pour toutes les observations du fichier `MontantReclFr_valid.csv`. Est-ce que vos prévisions sont bonnes ? Interprétez.

En utilisant toutes les observations du fichier `MontantReclFr_test.csv`, calculez la proportion des montants `mont_recl` qui se trouvent dans l'intervalle de confiance sur la prévision au niveau 95 %. Calculez aussi la racine carrée de l'erreur quadratique moyenne de prévision. Est-ce que votre modèle a une bonne performance ?

5% pour ce point.

Vous devez également fournir la sévérité prédite pour les polices contenues dans le fichier `MontantReclFr_test.csv`. Exportez vos prévisions, dans l'ordre, dans un fichier csv contenant deux colonnes (ID et prévision du montant de réclamation, en dollars), avec une commande du type

```
write.csv2(data.frame(ID=1:length(prevision), prevision),
"PrevisionEquipeXX.csv", row.names = FALSE)
```

Il n'y a rien à mentionner dans le rapport concernant le fichier `MontantReclFr_test.csv` et les prévisions reliées. Vous obtiendrez une note sur 5 pour la performance de votre modèle selon l'erreur quadratique moyenne de prévision calculée avec les vraies observations de `mont_recl` que votre superviseure ne vous a pas données.

Recommandations : Pour conclure votre rapport, donnez vos recommandations à votre superviseure (maximum 1/2 page). Le modèle que vous proposez est-il meilleur que celui qui est présentement utilisé ? Y a-t-il d'autres considérations ou ajustements dont vous devriez tenir compte ?

Annexe A : Listez les erreurs dans les données et les modifications apportées (maximum 1/2 page).

Les erreurs doivent être déjà corrigées dans l'analyse univariée.

Annexe B : Si vous avez effectué une transformation sur une ou des variables, expliquez pourquoi et quelle méthode vous avez utilisée (maximum une page).

Annexe C : Expliquez comment vous avez sélectionné les variables (maximum deux pages). Si vous avez fait des tests F partiels, ou une sélection selon un critère en particulier (critère AIC, Cp de Mallows), présentez les dans cette section. Votre superviseure doit pouvoir comprendre pourquoi certaines variables sont exclues du modèle alors que d'autres sont conservées.

Annexe D : Reproduisez le résumé du modèle choisi fourni par la commande `summary` en R.

4 Liste de vérification

Cette liste peut vous aider à vérifier votre travail avant la remise.

1. Réponses à toutes les questions listées dans la section 3 de cet énoncé
2. Orthographe et grammaire
3. Page titre avec le numéro d'équipe
4. Faits saillants (maximum 250 mots)
5. Table des matières
6. Pagination
7. Titres des tableaux et graphiques clairs
8. Les tableaux et graphiques sont tous mentionnés dans le texte
9. Les légendes et les titres des axes des graphiques sont clairs
10. La police dans les tableaux et graphiques est lisible (sans zoom)
11. Les commentaires sont clairs et concis
12. La notation est bien définie
13. Les calculs sont exacts
14. Les formules sont claires et correctes
15. Analyse univariée
16. Correction des erreurs dans les données, s'il y a lieu
17. Explication du modèle proposé
18. Vérification des postulats et de la multicolinéarité
19. Test de validité globale de la régression
20. Prévisions et intervalles de confiance pour les premières lignes du fichier `MontantReclFr_valid.csv`.
21. Prévisions pour le fichier `MontantReclFr_test.csv`.
22. Recommandations
23. Annexe A
24. Annexe B
25. Annexe C
26. Annexe D
27. Le rapport en format pdf est remis dans la boîte de dépôt.

Amusez-vous bien !