Titre : Prédiction des charges médicales

Classes : 4ème DS | Module : statistique | Année universitaire : 23-24

Motivation

Afin qu'une compagnie d'assurance gagne de l'argent, elle doit collecter plus de primes annuelles qu'elle n'en dépense en soins médicaux pour ses bénéficiaires. En conséquence, les assureurs investissent beaucoup de temps et d'argent pour développer des modèles permettant de prévoir avec précision les dépenses médicales.
Les charges médicales sont difficiles à estimer car les affections les plus coûteuses sont rares et apparemment aléatoires. Néanmoins, certaines conditions sont plus répandues chez certains segments de la population. Par exemple, le cancer du poumon est plus probable chez les fumeurs que chez les non-fumeurs, et les maladies cardiaques peuvent être plus probables chez les personnes obèses.

Objectifs et méthodologie

L'objectif de ce projet est de construire un modèle de régression permettant de prédire avec précision les charges médicales d'une personne et de déterminer les variables les plus influentes à cet effet. Les techniques de modélisation utilisées sont :

1. Régression linéaire classique.
2. Régression pénalisée (crête & lasso).
3. Régression en composantes principales.

Livrables

Le groupe chargé par ce projet est demandé de préparer :

1. R script.
2. Présentation.

Présentation de données

Pour cette analyse, nous utiliserons un ensemble de données simulées contenant les charges médicales des patients aux États-Unis. Le fichier **insurance.csv** comprend 1338 exemples de bénéficiaires actuellement inscrits au plan d'assurance, avec des caractéristiques indiquant les caractéristiques du patient ainsi que le total des charges médicaux facturés au plan pour l'année civile. Les variables explicatives sont :

- age : l'âge du patient.
- sex : le sexe du preneur d'assurance, homme ou femme.
- bmi : indice de masse corporelle (IMC), qui donne une idée du surpoids ou de l'insuffisance pondérale d'une personne par rapport à sa taille.
- children : le nombre d'enfants/personnes à charge couverts par le régime d'assurance.
- smoker : situation de l'assureur (fumeur ou non).
- region : lieu de résidence de l'assureur.

Travail demandé

Partie 1 : Exploration de données

1. Pré-traitement :
   - Importation des données (Importation, présentation et résumé statistique)
   - Traitement des valeurs manquantes et aberrantes.
2. Analyse univariée :
   - Etude de la distribution (normalité) des variables quantitatives (graphique & test statistique).
   - Etude des modalités (valeurs possibles) des variables qualitatives.
3. Analyse bivariée :
   En justifiant le choix des tests utilisés :
   - Etudier la corrélation entre les variables quantitatives (deux à deux).
   - Etudier la dépendance de la variable cible par chacune des variables qualitatives.

   Proposer un test statistique permettant de :
   - Etudier la dépendance de la variable cible par les deux variables qualitatives (sexe et smoker).
   - Etudier l'interaction entre ces variables qualitatives explicatives.

Partie 2 : Modèles linéaires

1. Régression linéaire multiple :

   Nous nous plaçons ici dans le contexte d'un modèle reliant une variable dépendante avec un certain nombre de variables explicatives (prédicteurs). Le modèle le plus répandu est la régression linéaire multiple. L'ajustement de ce type de modèle se fait souvent par les moindres carrés ordinaires.
   - Construire un modèle linéaire nommé « **modele1** » reliant les variables explicatives influentes sur la variable cible en expliquant la sélection des variables.
   - Diagnostiquer graphiquement le modèle développé et évaluer sa performance.
2. Régression linéaire améliorée :

   Le modèle linéaire « **modele1** » supposait uniquement une relation linéaire entre les variables. Cependant, en pratique dans l'assurance, le scénario peut être assez compliqué. Il faut prendre en compte ces complications et apporter des modifications subtiles au modèle. Nous proposons trois modifications :

   Modification 1 : **Relation non linéaire entre l'âge et les charges :**

   Les charges d'assurance maladie ont tendance à augmenter de manière disproportionnée avec l'âge. Ainsi, au lieu d'une relation linéaire, nous considérons une relation d'ordre supérieur avec l'âge.

   $$charges = \beta_0 + \beta_1\, age + \beta_2\, age^2$$

   Cela implique que le modèle doit prendre en compte un coefficient supplémentaire β_2. Par conséquent, nous incluons une variable (feature) supplémentaire basée sur le carré de l'âge.

   Modification 2 : **Convertir une variable numérique en un indicateur binaire :**

   L'bmi peut avoir un impact nul sur les charges médicales pour les individus ayant un poids normal, mais il peut être fortement lié à des coûts plus élevés pour les personnes obèses (c'est-à-dire un bmi de 30 ou plus). Nous pouvons modéliser cette relation en créant une variable indicatrice binaire qui vaut 1 si l'bmi est d'au moins 30 et 0 sinon. Le bêta estimé pour cette variable binaire indiquerait alors l'impact net moyen sur les dépenses médicales des personnes ayant un bmi de 30 ou plus, par rapport à celles ayant un bmi inférieur à 30.

   Pour créer cette caractéristique, nous pouvons utiliser la fonction **ifelse()**, qui pour chaque élément d'un vecteur teste une condition spécifiée et renvoie une valeur selon que la condition est vraie ou fausse. Pour un bmi supérieur ou égal à 30, nous renverrons 1, sinon 0. Nous pouvons ensuite inclure la variable $bmi30$ dans notre modèle amélioré.

   Modification 3 : **Ajout d'effet d'interaction :**

   Le fumer et l'obésité peuvent avoir des effets néfastes séparément, mais il est raisonnable de supposer que leurs effets combinés peuvent être pires que la somme de chacun d'eux pris isolément. Lorsque deux variables ont un effet combiné, on parle d'interaction. Si nous soupçonnons que deux variables

interagissent, nous pouvons tester cette hypothèse en ajoutant leur interaction au modèle. Nous considérons donc une interaction de ces deux variables dans le modèle ($smoker * ibm30$).

- Construire un modèle linéaire nommé « modele2 » avec les trois considérations précédentes.
- Evaluer ce modèle et le comparer avec l'ancien modèle en justifiant le choix des métriques et des tests utilisés.

3. Régression pénalisée :

   La précision d'un estimateur se mesure par son erreur quadratique moyenne. Cette erreur se décompose en deux terme le biais au carré et la variance. Si la variance est élevée, alors l'erreur l'est aussi et le modèle aura de mauvaise qualité de prédiction. Notre objectif est de chercher un modèle ayant la meilleure qualité prédictive, alors on cherche à diminuer la variance de chaque paramètre.
   - Effectuer une étude bibliographique sur la régression linéaire pénalisée (principe, différents types, et leur mise en œuvre).
   - Etablir cette méthode pour estimer les paramètres (coefficients) de notre modèle avec les deux types (ridge & lasso) et interpréter les résultats trouvés.

Partie 3 : Analyse multidimensionnelle

L'analyse multivariée est un ensemble de techniques a pour objectif de découvrir la structure, éventuellement compliquée, d'un tableau à plusieurs dimensions et de traduire par une structure plus simple et qui la résume le mieux.

1. Effectuer une étude bibliographique sur l'analyse en composantes principales.
2. En se basant sur cette méthode, réduire la dimension de la base de données modifiée (base de données initiale plus les trois nouvelles variables construises).
3. Construire un modèle linéaire nommée « modele3 » reliant la variable cible avec les variables explicatives de la nouvelle base de données.
4. Faire une étude comparative entre les trois modèles développés.

Références

Ref 1 : B Lantz.Machine Learning with R, 2019.

Ref 2 : F Brun, É Doutart, F Duyme, M El Jabri, K Fauvel. Data science pour l'agriculture et l'environnement-Méthodes et applications avec R et Python, 2021.

Ref 3 : C Duby, S Robin. Analyse en composantes principales - Institut National Agronomique, Paris-Grignon, 2006.