Analyse de données
M1 Statistique et économétrie - 2012
V. Monbet

Projet : préparation du contrôle terminal

Toutes les questions doivent être étudiées en détails. Vous devez être capable de justifier la démarche et les choix méthodologiques, écrire les modèles, justifier vos réponses, commentez les résultats, etc.

Dans ce projet, il s'agit d'aider une organisation non gouvernementale à cibler les populations auprès desquelles elle doit intervenir pour améliorer l'usage de la contraception féminine en Indonésie. Le jeux de données dont nous disposons est un extrait d'un sondage réalisé en 1987 sur une population de femmes mariées. Les variables observées sont les suivantes

1. Age de la femme interrogée.
2. Niveau d'éducation de la femme (1=faible, 2, 3, 4=élevé)
3. Niveau d'éducation du mari (1=faible, 2, 3, 4=élevé)
4. Nombre d'enfants
5. Réligion (0=non Islam, 1=Islam)
6. Travaille actuellement (0=Oui, 1=Non)
7. Occupation du mari (1, 2, 3, 4)
8. Index de niveau de vie (1=faible, 2, 3, 4=élevé)
9. Exposition aux médias (0=Good, 1=Not good)
10. Méthode contraceptive utilisée (1=Aucune, 2=Long terme, 3=Court terme)

On cherche à prédire l'utilisation d'une contraception.

Les données `projet2012.data` sont disponibles sur la page `http://perso.univ-rennes1.fr/valerie.monbet/Cours_AD/AD.html`. Importez les données dans le logiciel que vous avez choisi d'utiliser. Pensez à définir les variables qualitatives comme des variables binaires ou ordinales.

1 Pré-traitements

1. Quelle est la nature des variables ?
2. Estimer et représenter la distribution de la variable *méthode de contraception utilisée*.
3. Le codage de certaines variables est contre-intuitif. Procéder à un recodage.
4. Les variables *age* et *nombre d'enfants* sont elles corrélées ?
5. Les variables *Niveau d'étude* et *Niveau d'étude du mari* sont-elles dépendantes ?

2 Analyse descriptive

1. Réaliser une analyse factorielle permettant de mettre en évidence des liens entre les différentes modalités des variables qualitatives. Vous pourrez choisir d'introduire certaines variables en variables supplémentaires. Il est notamment judicieux de garder la variable à prédire en variable supplémentaire car elle joue un rôle particulier.
2. Proposez une discrétisation des variables quantitatives. Vos choix méthodologiques devront être justifiés.
3. Reprendre l'analyse factorielle en incluant les variables obtenues après discrétisation.

3 Analyse discriminante

On va maintenant construire des règles de décision permettant de discriminer la population en fonction de l'utilisation d'une méthode contraceptive.

3.1 Préparation des données

1. Générer aléatoirement un échantillon d'apprentissage et un échantillon de test. Le premier sera utilisé pour calibrer les modèles et le second pour les valider ; ceci pour tous les modèles que vous construisez. Vous pouvez utiliser les fonctions `sample` et `setdiff` de `R` ou la procédure `PROC SURVEY` de `SAS`
2. Assurez vous que la distribution de la variable cible dans l'échantillon de test est similaire à la distribution de cette variable dans la population interrogée.

3.2 Modélisation pour la variable cible à 3 modalités

1. **Modèle de Bayes naïf**
   Proposez un modèle non paramétrique basé sur les plus proches voisins. Comment choisissez-vous le nombre optimal de voisins ? Donner la table de contingence. Estimez le risque de Bayes.

2. **Analyse discriminante de Fisher**
   Ici, nous pouvons utiliser l'analyse discriminante de Fisher si nous combinons une analyse factorielle des correspondances multiples (ACM) et une analyse discriminante.
   (a) Estimer les composantes principales d'une ACM sur les variables qualitatives explicatives. Ne retenir que les composantes associées à une inertie assez grande.
   (b) Faire une analyse de la variance multivariée (manova) pour vérifier que l'analyse discriminante a un sens.
   (c) Construire une règle de décision linéaire avec pour variables explicatives les composantes principales retenues ainsi que les variables quantitatives observées. Vous pouvez utiliser la fonction `lda` sous `R` ou la procédure `PROC DISCRIM` sous `SAS`. Donner la table de contingence. Estimer le risque de Bayes.
3. Comparer et critiquer les deux modèles obtenus.

3.3 Modélisation pour la variable cible à 2 modalités

La variable *méthode contraceptive utilisée* semble être difficile à prédire. Comme nous cherchons essentiellement à cibler les femmes n'utilisant pas de contraceptif, nous allons poser le problème autrement.

1. Créer une nouvelle variable *"utilise une méthode contraceptive (OUI ou NON)"* à partir de la variable observée *"méthode contraceptive utilisée"*. Cette nouvelle variable devient la variable à prédire.
2. **Régression logistique** pour prédire la variable *utilise une méthode contraceptive.* Ici vous pouvez travailler directement avec les variables d'origine.
   (a) Proposer un modèle de régression logistique pour prédire la variable *utilise une méthode contraceptive*. Sous `R` on utilise la fonction `glm` et sous `SAS` la procédure `PROC LOGISTIC`. Utiliser les tests statistiques associés à la régression logistique pour sélectionner les variables discriminantes (par exemple en utilisant la fonction `step` [1] de `R` ou l'option `stepwise` de la `PROC LOGISTIC`). Pensez à éventuellement introduire des interactions.
   (b) Lister les variables que vous conservez. Et écrire le modèle.
   (c) Donner la table de contingence (pour un seuil de 0.5) et estimer le risque de Bayes.
3. **Analyse discriminante de Fisher** pour prédire la variable *utilise une méthode contraceptive.*

[1] Voir un exemple sous le lien http ://data.princeton.edu/R/glms.html

(a) Proposer un modèle basé sur l'analyse discriminante de Fisher. Justifiez le choix du modèle. Vous écrirez explicitement la règle de décision, en définissant correctement vos notations.
Notez qu'en ACM, il est possible de considérer des individus supplémentaires et ainsi d'obtenir leurs coordonnées même s'ils ne participent pas à l'analyse.

(b) Donner la table de contingence et estimer le risque de Bayes.

4. D'après les questions précédentes, pouvez-vous expliquer quel est le profil type des femmes qui n'utilisent pas de contraception ? Justifier les réponses.