ENSAE-ENSAI Formation continue (Cepe)

Certificat
Chargé d'études statistiques

# Fiche méthode : Régression logistique dichotomique

Martin Chevalier (Insee)

## 1 Quand utiliser une régression logistique dichotomique ?

— **une caractéristique** que l'on souhaite expliquer par **plusieurs autres**
**Exemple** Chômage en fonction de l'âge, du sexe et du diplôme

— caractéristique à expliquer **binaire** (2 modalités exactement)
**Exemple** Réussite à un examen, décision d'achat

## 2 Quel est l'objectif de la modélisation ?

— prédiction : **reconstituer au mieux** la variable expliquée à partir des seules variables explicatives
**Exemple** Détection des *spams* dans un service de messagerie électronique

— interprétation : mettre en évidence des **relations statistiquement significatives** entre certaines variables explicatives et la variable expliquée
**Exemple** Analyse des déterminants de la poursuite d'études supérieures

## 3 Avant de commencer

— analyse univariée de la variable expliquée et des variables explicatives : état des lieux des **valeurs manquantes**, détection de **valeurs atypiques**, **recodages** eventuels

— analyse bivariée de la variable expliquée par les variables explicatives : mise en évidence des relations les plus déterminantes et d'éventuels **effets de structure**

— modèle à finalité interprétative : quelles sont les **hypothèses de travail** ?
**Exemple** Plus le niveau de diplôme des parents est élevé, plus l'enfant a une probabilité importante de poursuivre ses études dans l'enseignement supérieur

## 4 Construire le modèle

— modèle à finalité prédictive : maximisation de la capacité prédictive du modèle (pourcentage de concordance, courbe ROC), sélection (semi-)automatique de variables, validation du modèle sur un échantillon de test

— modèle à finalité interprétative : ajout progressif des variables en lien avec les hypothèses de travail, analyse des tests associés aux effets de type III, introduction de variables croisées

— dichotomisation des variables explicatives qualitatives, manuellement ou avec l'instruction `CLASS` (accompagnée de l'option `PARAM = REF`) :

```
PROC LOGISTIC DATA = enquete_emploi;
    CLASS sexe (REF = "Homme") diplome (REF = "Bac") / PARAM = REF;
    MODEL chomage = age sexe age * sexe diplome;
RUN;
```

## 5 Finalité prédictive : Utiliser les prédictions du modèle

— modèle estimé et validé sur un **échantillon d'apprentissage** pour lequel la variable expliquée et les variables explicatives sont connues

— utilisation des coefficients pour **prédire sur un nouvel échantillon** la valeur de la variable expliquée à partir des seules variables explicatives

## 6 Finalité interprétative : Interpréter les résultats du modèle

— signe, amplitude et significativité des **coefficients** : erreur-standard, statistique de test, p-valeur, intervalle de confiance

— cas particulier des **variables qualitatives** : interprétation par rapport à la modalité de référence, choix de la modalité de référence

— ***odds-ratios*** : interprétation en termes de rapport de chance, inférence à partir de l'intervalle de confiance

— **effets marginaux moyens** : calcul par différentes méthodes (document de travail de l'Insee, aide de SAS), interprétation en termes d'écart en points de pourcentage

— **tests d'hypothèses complexes** : test d'hypothèses en lien avec la problématique de l'étude, en particulier en présence de variables d'interaction

## 7 Finalité interprétative : Présenter les résultats du modèle

— **Adapter la présentation au public visé** : coefficients, *odds-ratios* ou effets marginaux moyens, signes + et -, graphiques représentant visuellement les *odds-ratio* et leur intervalle de confiance

— **Faciliter la compréhension** : affichage du libellé plutôt que du nom des variables, ajout de la modalité de référence pour les variables qualitatives, utilisation de l'ordre naturel pour les variables ordonnées

— Mettre en avant les résultats **statistiquement significatifs** : étoiles (* : 10 %, ** : 5 %, *** : 1 %), passage en gras, mise en « ns » ou à blanc des valeurs non-significatives

— Accompagner le tableau de résultat d'**éléments de contexte** : titre, numéro de tableau (et appel dans le texte), source de données utilisées et période de validité, champ couvert, note de lecture