Régression logistique

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Professional Rodeo Cowboys Association

L'exemple suivant est inspiré de l'article

> *Daneshvary, R. et Schwer, R. K. (2000) The Association Endorsement and Consumers' Intention to Purchase. Journal of Consumer Marketing 17, 203-213.*

Objectif: Les auteurs cherchent à voir si le fait qu'un produit soit recommandé par le *Professional Rodeo Cowboys Association* (PRCA) a un effet sur les intentions d'achats.

Données du PRCA

On dispose de 500 observations sur les variables suivantes dans la base de données `logit1`:

- Y: seriez-vous intéressé à acheter un produit recommandé par le PRCA
  - 0: non
  - 1: oui
- X_1: quel genre d'emploi occupez-vous?
  - 1: à la maison
  - 2: employé
  - 3: ventes/services
  - 4: professionnel
  - 5: agriculture/ferme
- X_2: revenu familial annuel
  - 1: moins de 25 000
  - 2: 25 000 à 39 999
  - 3: 40 000 à 59 999
  - 4: 60 000 à 79 999
  - 5: 80 000 et plus

Données du PRCA

- X_3: sexe
  - 0: homme
  - 1: femme
- X_4: avez-vous déjà fréquenté une université?
  - 0: non
  - 1: oui
- X_5: âge (en années)
- X_6: combien de fois avez-vous assisté à un rodéo au cours de la dernière année?
  - 1: 10 fois ou plus
  - 2: entre six et neuf fois
  - 3: cinq fois ou moins

Régression logistique

Expliquer le comportement de la **moyenne** d'une variable binaire $Y \in \{0, 1\}$ en utilisant un modèle de régression avec p variables explicatives $\mathsf{X}_1, \dots, \mathsf{X}_p$.

$$\underbrace{\mathsf{E}(Y = 1 \mid \mathbf{X})}_{\text{moyenne théorique}} = \underbrace{\mathsf{Pr}(Y = 1 \mid \mathbf{X})}_{\text{probabilité de succès}} = p$$

Objectif de la régression

1) **Inférence** : comprendre comment et dans quelles mesures les variables **X** influencent la probabilité que $Y = 1$.
2) **Prédiction** : développer un modèle pour prévoir des valeurs de Y ou la probabilité de succès à partir des **X**.

Exemples

- Est-ce qu'un client potentiel va répondre favorablement à une offre promotionnelle?
- Est-ce qu'un client est satisfait du service après-vente?
- Est-ce qu'un client va faire faillite ou non au cours des trois prochaines années.

Modéliser une probabilité avec une régression linéaire?

- Aucune contrainte sur le prédicteur linéaire $\beta_0 + \beta_1 \mathsf{X}_1 + \cdots + \beta_p \mathsf{X}_p$
  - retourne des probabilités négatives ou supérieures à 1!
- les données binaires ne respectent pas le postulat d'égalité des variances
  - invalide résultat des tests d'hypothèse sur coefficients.

Illustration: linéaire vs logistique

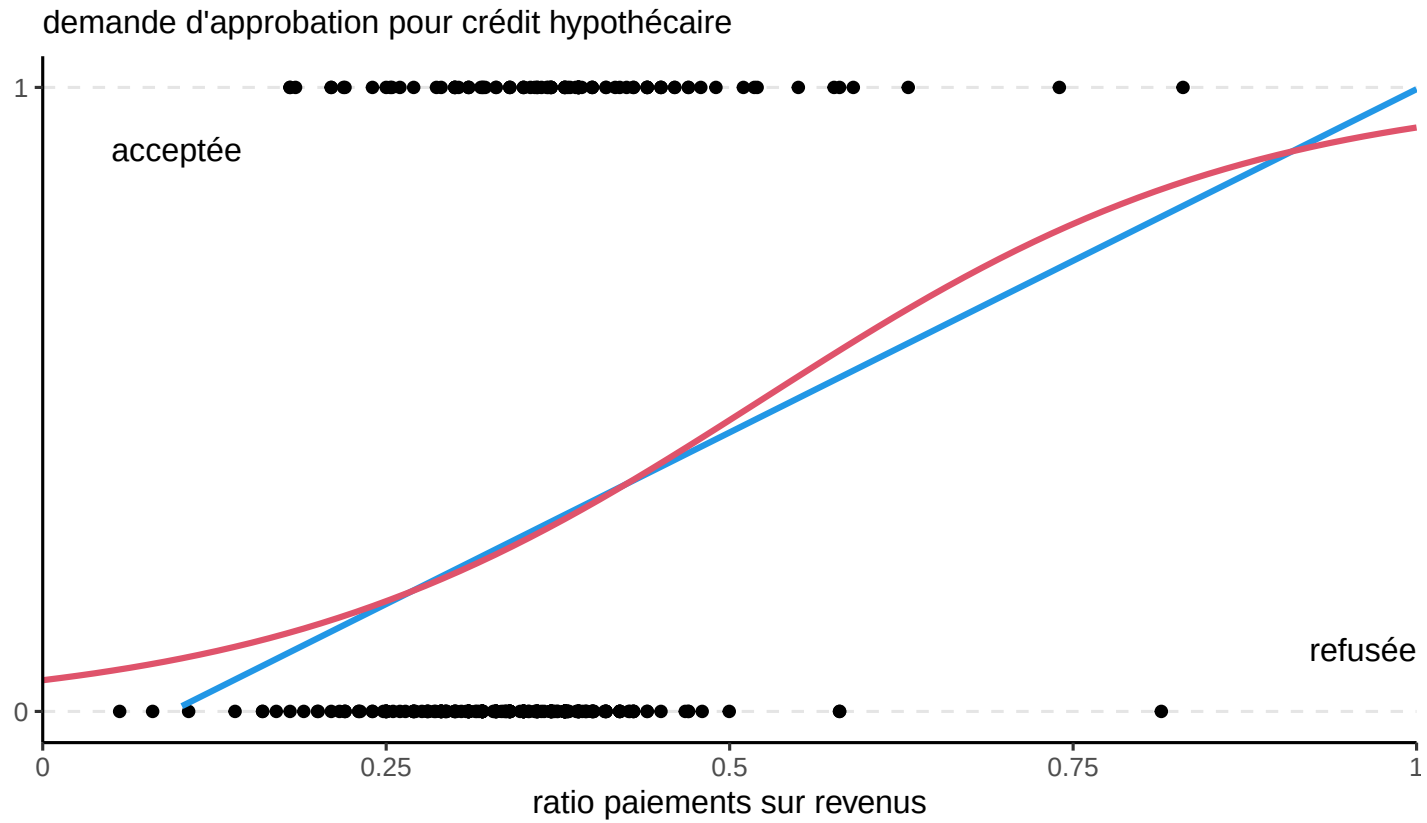

Figure 1: Données de la réserve de Boston sur l'approbation de prêts hypothécaires (1990); données tirées de Stock et Watson (2007).

Fonction de liaison

Idée: appliquer une transformation au **prédicteur linéaire**

$$\eta = \beta_0 + \beta_1 \mathsf{X}_1 + \cdots + \beta_p \mathsf{X}_p$$

pour que la prédiction soit entre zéro et un.

On considère

$$p = \mathsf{expit}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{1}{1 + \exp(-\eta)}.$$

Courbe sigmoïde

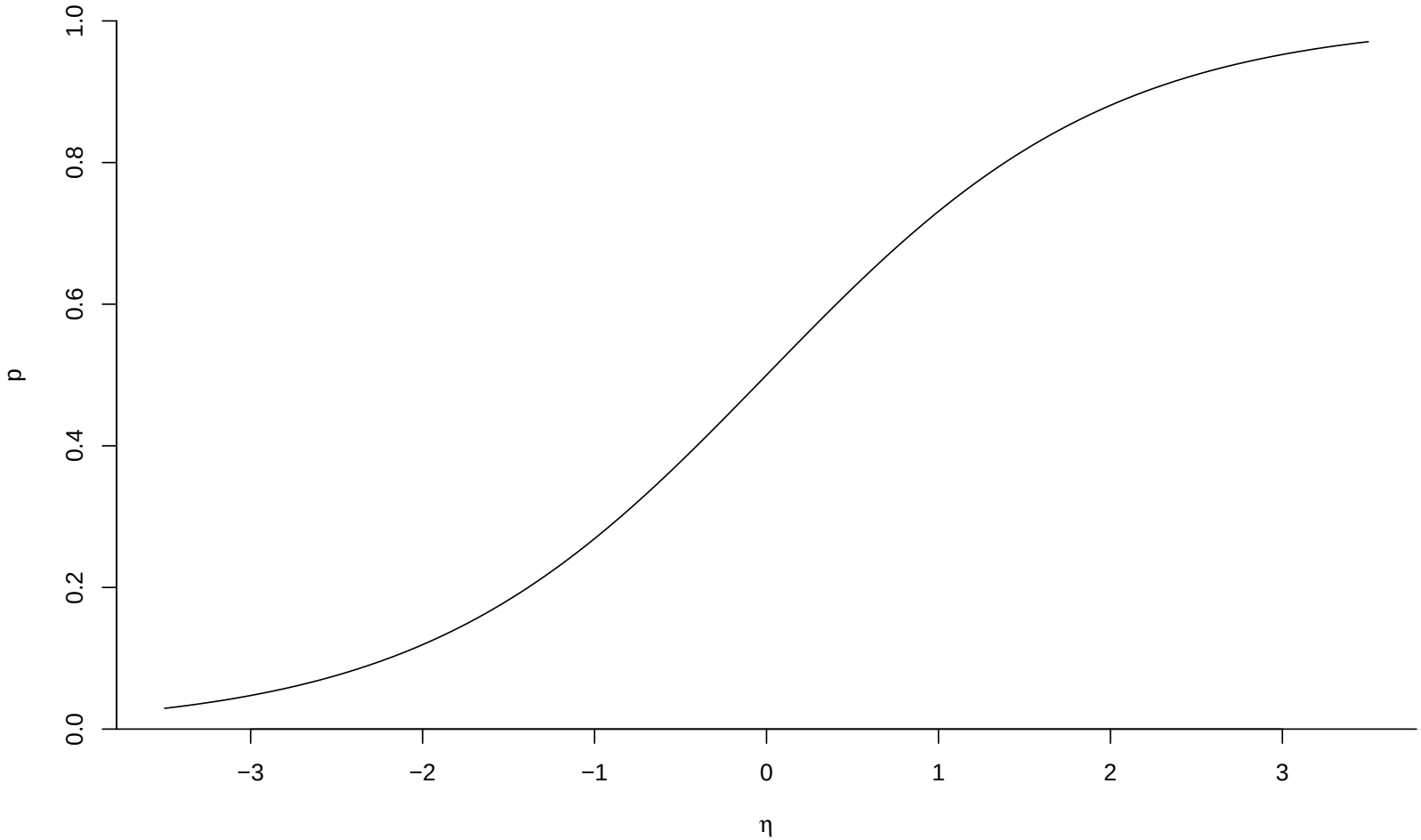

Figure 2: Valeurs ajustées du modèle de régression logistique en fonction du prédicteur linéaire η.

Cote

La cote donne le ratio de la probabilité de succès ($Y = 1$) sur la probabilité d'échec ($Y = 0$).

$$\text{cote}(p) = \frac{p}{1-p} = \frac{\Pr(Y = 1 \mid \mathbf{X})}{\Pr(Y = 0 \mid \mathbf{X})}.$$