MATH 60604
Modélisation statistique
§ 4a - Modèles linéaires généralisés

Léo Belzile

HEC Montréal
Département de sciences de la décision

Introduction

- Les modèles linéaires ne sont adéquats que pour les variables réponses qui suivent (conditionnellement) une loi normale.
- Dans plusieurs contextes, les variables réponse Y à disposition sont
  - binaires,
  - entières, c'est-à-dire des variables de dénombrement,
  - continues, mais non-négatives,
- On considère des lois adéquates pour des données binaires, des proportions et des variables de dénombrement, afin de faire de l'inférence basée sur la vraisemblance.

Variables réponses binaires

- Si la variable réponse Y vaut soit 0, soit 1, on peut postuler une loi Bernoulli pour Y, soit

$$\mathsf{P}\left(Y=y\right)=\pi^{y}(1-\pi)^{1-y},\quad y=0,1.$$

  Il en découle que $\mathsf{E}\left(Y\right)=\pi$ et $\mathsf{Var}\left(Y\right)=\pi(1-\pi)$.
- Par convention, les zéros représentent des échecs (non) et les uns des réussites (oui).
- Exemples de questions de recherches comprenant une variable réponse binaire:
  - est-ce qu'un client potentiel a répondu favorablement à une offre promotionnelle?
  - est-ce qu'un client est satisfait du service après-vente?
  - est-ce qu'une firme va faire faillite au cours des trois prochaines années?
  - est-ce qu'un participant à une étude réussit une tâche?

Variables réponses binaires cumulées

Si les données représentent la somme d'événements Bernoulli indépendants, la loi du nombre de réussites Y sur le nombre d'essais m est binomiale avec fonction de masse

$$\mathrm{P}\left(Y = y\right) = \binom{m}{y}\pi^{y}(1-\pi)^{m-y}, \quad y = 0, 1.$$

La vraisemblance pour un échantillon $\mathrm{Bin}(m, \pi)$ est (à constante de normalisation près qui ne dépend pas de π) la même que pour un échantillon aléatoire de m variables Bernoulli indépendantes. L'espérance est $\mathrm{E}\left(Y\right) = m\pi$ et la variance $\mathrm{Var}\left(Y\right) = m\pi(1-\pi)$.

Variables de dénombrement

- Si la probabilité d'un événement est **rare**, on suppose souvent que Y, le nombre de réussites dans un intervalle de temps donné, suit une loi de Poisson,

$$\mathsf{P}\left(Y = y\right) = \frac{\exp(-\mu)\mu^y}{y!}, \quad y = 0, 1, 2, \ldots$$

- Le paramètre μ de la loi de Poisson est à la fois l'espérance et la variance de la variable, $\mathsf{E}\left(Y\right) = \mathsf{Var}\left(Y\right) = \mu$.
- Exemples de questions de recherches comprenant une variable réponse de dénombrement:
  - nombre de réclamations faites par un client d'une compagnie d'assurance au cours d'une année.
  - nombre d'achats effectués par un client depuis un mois.
  - nombre de tâches réussies par un participant lors d'une étude.

Notation pour les modèles linéaires généralisés

- Le point de départ est le même que pour la régression linéaire:
  - On dispose d'un échantillon d'observations indépendantes

$$(Y_i, \mathrm{X}_{i1}, \ldots, \mathrm{X}_{ip}), \quad i = 1, \ldots, n$$

    où Y est la variable réponse et $\mathrm{X}_1, \ldots, \mathrm{X}_p$ des variables explicatives supposées connues et fixes (non-aléatoires).
  - Le but est modéliser la variable réponse en fonction des variables explicatives.
- On dénote μ_i la moyenne (conditionnelle) de Y_i étant données les variables explicatives,

$$\mu_i = \mathrm{E}\left(Y_i \mid \mathrm{X}_{i1}, \ldots, \mathrm{X}_{ip}\right).$$

- On dénote par η_i la combinaison linéaire des variables explicatives qui servira à modéliser la variable réponse,

$$\eta_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}.$$

Composantes du modèle linéaire généralisé

Trois composantes sont nécessaires pour définir un modèle linéaire généralisé

- Une loi de probabilité pour la variable réponse Y qui fait partie de la famille exponentielle (normale, binomiale, Poisson, gamma, ...).
- Un prédicteur linéaire $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.
- Une fonction monotone g, appelée fonction de liaison, que relie la moyenne de Y_i aux variables explicatives, $g(\mu_i) = \eta_i$, d'où

$$\begin{aligned} & g(\mu_i) = \eta_i = \beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip} \\ \Leftrightarrow \quad & \mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip}). \end{aligned}$$

Fonction de liaison

- Dans le modèle de régression linéaire ordinaire, on n'impose pas de contraintes aux valeurs prises par la moyenne μ_i et $\widehat{\mu}_i = \widehat{\eta}_i$ peut prendre des valeurs arbitraires dans l'intervalle $(-\infty, \infty)$.
- En revanche, les moyennes de certaines variables réponses sont contraintes
  - variables Bernoulli/binomiales: la moyenne $\mu = \pi$ doit être dans l'intervalle $(0, 1)$.
  - variables Poisson: la moyenne μ doit être positive.
- Un choix adéquat de fonction de liaison pour μ_i permet une transformation de la combinaison linéaire η_i de telle sorte qu'aucune contrainte numérique n'est imposée sur les paramètres $\boldsymbol{\beta}$.

Choix de la fonction de liaison

Certains choix de fonctions de liaisons facilitent l'interprétation des paramètres ou l'optimisation avec la fonction de vraisemblance.

- Pour les lois Bernoulli et binomiale, la fonction de liaison la plus utilisée est la fonction logit,

$$\mathrm{logit}(\mu) := \ln\left(\frac{\mu}{1-\mu}\right) = \eta \quad \Leftrightarrow \quad \mu = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

- Pour la loi Poisson, la fonction de liaison canonique est le logarithme naturel,

$$\ln(\mu) = \eta \quad \Leftrightarrow \quad \mu = \exp(\eta).$$

- Pour la loi normale, la fonction de liaison est la fonction identité, donc $\mu = \eta$.

Cas particulier de la régression linéaire

- La régression linéaire ordinaire est un cas spécial de régression linéaire généralisée, avec

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip} + \varepsilon_i, \qquad (i = 1, \ldots, n)$$

  où $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{No}(0, \sigma^2)$, c'est-à-dire que $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables indépendantes et identiquement distribuées de loi normale avec moyenne 0 et variance σ^2.
- De façon équivalente,

$$Y_i \mid \mathbf{X}_i \stackrel{\text{ind}}{\sim} \mathsf{No}(\beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip}, \sigma^2)$$

- On déduit que la régression linéaire est un modèle linéaire généralisé avec
  - une loi normale pour la réponse et
  - la fonction identité comme fonction de liaison.