La régression logistique

Inès Vati, Clara Sicard-Benmedjahed, Inès Larroche

Statistiques et analyse de données

Janvier 2022

Plan de l'exposé

Cadre général

- Variable aléatoire Y, $Y \in \{0, 1\}$
- Y dépend d'un paramètre $x \in \mathbb{R}^p$, $x = (x_1, ..., x_p)$
- $\mathbb{P}(Y = 0) = p(0|x)$ et $\mathbb{P}(Y = 0) = p(1|x)$
- $p(1|x) = \phi(\beta_0 + \sum_{i=1}^{p}(\beta_i x^i))$
- $\phi(t) = \frac{1}{1+\exp -t} \in [0, 1]$, la loi logistique

Cadre général

- Variable aléatoire Y, $Y \in \{0, 1\}$
- Y dépend d'un paramètre $x \in \mathbb{R}^p$, $x = (x_1, ..., x_p)$
- $\mathbb{P}(Y = 0) = p(0|x)$ et $\mathbb{P}(Y = 0) = p(1|x)$
- $p(1|x) = \phi(\beta_0 + \sum_{i=1}^{p}(\beta_i x^i))$
- $\phi(t) = \frac{1}{1+\exp -t} \in [0, 1]$, la loi logistique

Cadre général

- Variable aléatoire Y, $Y \in \{0, 1\}$
- Y dépend d'un paramètre $x \in \mathbb{R}^p$, $x = (x_1, ..., x_p)$
- $\mathbb{P}(Y = 0) = p(0|x)$ et $\mathbb{P}(Y = 0) = p(1|x)$
- $p(1|x) = \phi(\beta_0 + \sum_{i=1}^{p}(\beta_i x^i))$
- $\phi(t) = \frac{1}{1+\exp -t} \in [0, 1]$, la loi logistique

Cadre général

- Variable aléatoire Y, $Y \in \{0,1\}$
- Y dépend d'un paramètre $x \in \mathbb{R}^p$, $x = (x_1, ..., x_p)$
- $\mathbb{P}(Y = 0) = p(0|x)$ et $\mathbb{P}(Y = 0) = p(1|x)$
- $p(1|x) = \phi(\beta_0 + \sum_{i=1}^{p}(\beta_i x^i))$
- $\phi(t) = \frac{1}{1+\exp -t} \in [0,1]$, la loi logistique

Cadre général

- Variable aléatoire Y, $Y \in \{0, 1\}$
- Y dépend d'un paramètre $x \in \mathbb{R}^p$, $x = (x_1, ..., x_p)$
- $\mathbb{P}(Y = 0) = p(0|x)$ et $\mathbb{P}(Y = 0) = p(1|x)$
- $p(1|x) = \phi(\beta_0 + \sum_{i=1}^{p}(\beta_i x^i))$
- $\phi(t) = \frac{1}{1+\exp -t} \in [0, 1]$, la loi logistique

Modèle

Estimation avec le maximum de vraisemblance

- On estime chaque β_j avec la méthode du maximum de vraisemblance
- Variables de Bernouilli, indépendantes $(y_1, ..., y_n)$
- $\forall i \in [1, n], y_i$ suit une $\mathcal{B}(p(1|x_i))$
- Calcul de $L_n(y_n, x_n, \beta)$
- $\forall j \in [1, p], \frac{\partial l_n}{\beta_j} = 0 => \hat{\beta}_j$
- Estimateur de la loi Y :

$$\hat{y_n} = \begin{cases} 1 \text{ si } \hat{p}(1|x) \geq p_0 \\ 0 \text{ sinon} \end{cases}$$

Modèle

Estimation avec le maximum de vraisemblance

- On estime chaque β_j avec la méthode du maximum de vraisemblance
- Variables de Bernouilli, indépendantes $(y_1, ..., y_n)$
- $\forall i \in [1, n], y_i$ suit une $\mathcal{B}(p(1|x_i))$
- Calcul de $L_n(y_n, x_n, \beta)$
- $\forall j \in [1, p], \frac{\partial l_n}{\beta_j} = 0 => \hat{\beta}_j$
- Estimateur de la loi Y :

$$\hat{y_n} = \begin{cases} 1 \text{ si } \hat{p}(1|x) \geq p_0 \\ 0 \text{ sinon} \end{cases}$$

Modèle

Estimation avec le maximum de vraisemblance

- On estime chaque β_j avec la méthode du maximum de vraisemblance
- Variables de Bernouilli, indépendantes $(y_1, ..., y_n)$
- $\forall i \in [1, n], y_i$ suit une $\mathcal{B}(p(1|x_i))$
- Calcul de $L_n(y_n, x_n, \beta)$
- $\forall j \in [1, p], \frac{\partial l_n}{\beta_j} = 0 => \hat{\beta}_j$
- Estimateur de la loi Y :

$$\hat{y_n} = \begin{cases} 1 \text{ si } \hat{p}(1|x) \geq p_0 \\ 0 \text{ sinon} \end{cases}$$

Modèle

Estimation avec le maximum de vraisemblance

- On estime chaque β_j avec la méthode du maximum de vraisemblance
- Variables de Bernouilli, indépendantes $(y_1, ..., y_n)$
- $\forall i \in [1, n], y_i$ suit une $\mathcal{B}(p(1|x_i))$
- Calcul de $L_n(y_n, x_n, \beta)$
- $\forall j \in [1, p], \frac{\partial l_n}{\beta_j} = 0 => \hat{\beta}_j$
- Estimateur de la loi Y :

$$\hat{y_n} = \begin{cases} 1 \text{ si } \hat{p}(1|x) \geq p_0 \\ 0 \text{ sinon} \end{cases}$$

Modèle

Estimation avec le maximum de vraisemblance

- On estime chaque β_j avec la méthode du maximum de vraisemblance
- Variables de Bernouilli, indépendantes $(y_1, ..., y_n)$
- $\forall i \in [1, n], y_i$ suit une $\mathcal{B}(p(1|x_i))$
- Calcul de $L_n(y_n, x_n, \beta)$
- $\forall j \in [1, p], \frac{\partial l_n}{\beta_j} = 0 => \hat{\beta}_j$
- Estimateur de la loi Y :

$$\hat{y_n} = \begin{cases} 1 \text{ si } \hat{p}(1|x) \geq p_0 \\ 0 \text{ sinon} \end{cases}$$

Modèle

Estimation avec le maximum de vraisemblance

- On estime chaque β_j avec la méthode du maximum de vraisemblance
- Variables de Bernouilli, indépendantes $(y_1, ..., y_n)$
- $\forall i \in [1, n], y_i$ suit une $\mathcal{B}(p(1|x_i))$
- Calcul de $L_n(y_n, x_n, \beta)$
- $\forall j \in [1, p], \frac{\partial l_n}{\beta_j} = 0 => \hat{\beta}_j$
- Estimateur de la loi Y :

$$\hat{y_n} = \begin{cases} 1 \text{ si } \hat{p}(1|x) \geq p_0 \\ 0 \text{ sinon} \end{cases}$$

Modèle

Likelihood Ratio test

- Y dépend t-elle vraiment du paramètre x ?
- Hypothèse nulle (cas p=1) $H_0 = \{\beta_1 = 0\}$
- Hypothèse alternative $H_0 = \{\beta_1 \neq 0\}$
- Statistique de test : $\Lambda(y_n|x_n) = 2log(\frac{\text{MLE model}}{}$
- Calcul de la p-valeur

Modèle

Likelihood Ratio test

- Y dépend t-elle vraiment du paramètre x ?
- Hypothèse nulle (cas p=1) $H_0 = \{\beta_1 = 0\}$
- Hypothèse alternative $H_0 = \{\beta_1 \neq 0\}$
- Statistique de test : $\Lambda(y_n|x_n) = 2log(\frac{\text{MLE model}}{}$
- Calcul de la p-valeur

Modèle

Likelihood Ratio test

- Y dépend t-elle vraiment du paramètre x ?
- Hypothèse nulle (cas p=1) $H_0 = \{\beta_1 = 0\}$
- Hypothèse alternative $H_0 = \{\beta_1 \neq 0\}$
- Statistique de test : $\Lambda(y_n|x_n) = 2log(\frac{\text{MLE model}}{}$
- Calcul de la p-valeur

Modèle

Likelihood Ratio test

- Y dépend t-elle vraiment du paramètre x ?
- Hypothèse nulle (cas p=1) $H_0 = \{\beta_1 = 0\}$
- Hypothèse alternative $H_0 = \{\beta_1 \neq 0\}$
- Statistique de test : $\Lambda(y_n|x_n) = 2log(\frac{\text{MLE model}}{}$
- Calcul de la p-valeur

Modèle

Likelihood Ratio test

- Y dépend t-elle vraiment du paramètre x ?
- Hypothèse nulle (cas p=1) $H_0 = \{\beta_1 = 0\}$
- Hypothèse alternative $H_0 = \{\beta_1 \neq 0\}$
- Statistique de test : $\Lambda(y_n|x_n) = 2log(\frac{\text{MLE model}}{}$
- Calcul de la p-valeur

Analyse d'un jeu de données

Position du problème

- Sujet : Détection du cancer du pancréas pour un ensemble de patients donné. Analyse de l'influence des biomarkers REG1B, TFF1, REG1A, présents dans l'urine.
- Y= variable aléatoire détectant la présence du cancer.
- Le patient est atteint :
  $\mathbb{P}(Y = 1) = p(1|x)\text{avec}x = (REG1B, TFF1, REG1A)$
- Provenance des données : Indiana School of Medecine, USA

Analyse d'un jeu de données

Position du problème

- Sujet : Détection du cancer du pancréas pour un ensemble de patients donné. Analyse de l'influence des biomarkers REG1B, TFF1, REG1A, présents dans l'urine.
- Y= variable aléatoire détectant la présence du cancer.
- Le patient est atteint :
  $\mathbb{P}(Y = 1) = p(1|x)$avec$x = (REG1B, TFF1, REG1A)$
- Provenance des données : Indiana School of Medecine, USA

Analyse d'un jeu de données

Position du problème

- Sujet : Détection du cancer du pancréas pour un ensemble de patients donné. Analyse de l'influence des biomarkers REG1B, TFF1, REG1A, présents dans l'urine.
- Y= variable aléatoire détectant la présence du cancer.
- Le patient est atteint :
  $\mathbb{P}(Y = 1) = p(1|x) \text{avec} x = (REG1B, TFF1, REG1A)$
- Provenance des données : Indiana School of Medecine, USA

Analyse d'un jeu de données

Position du problème

- Sujet : Détection du cancer du pancréas pour un ensemble de patients donné. Analyse de l'influence des biomarkers REG1B, TFF1, REG1A, présents dans l'urine.
- Y= variable aléatoire détectant la présence du cancer.
- Le patient est atteint : $\mathbb{P}(Y=1) = p(1|x) \text{avec} x = (REG1B, TFF1, REG1A)$
- Provenance des données : Indiana School of Medecine, USA

Analyse d'un jeu de données

Méthode

- Division du dataset en un dataset d'entraînement et un dataset de validation
- Pré-traitement des données
- Calcul des paramètres avec le modèle de la régression logistique.

Analyse d'un jeu de données

Méthode

- Division du dataset en un dataset d'entraînement et un dataset de validation
- Pré-traitement des données
- Calcul des paramètres avec le modèle de la régression logistique.

Analyse d'un jeu de données

Méthode

- Division du dataset en un dataset d'entraînement et un dataset de validation
- Pré-traitement des données
- Calcul des paramètres avec le modèle de la régression logistique.

Analyse d'un jeu de données

Méthode

- Division du dataset en un dataset d'entraînement et un dataset de validation
- Pré-traitement des données
- Calcul des paramètres avec le modèle de la régression logistique.

Analyse d'un jeu de données

Résultats

- Valeur des paramètres
- Limites, améliorations.

Analyse d'un jeu de données

Résultats

- Valeur des paramètres
- Limites, améliorations.

Conclusion