M1 SEP-MA CS-MA SEP-APE
2020 - 2021
SEP0832-Méthodes d'échantillonnage
IE, 15 Avril 2021
Durée : 1h00

Voici d'autres **consignes** à respecter en plus de celles présentées dans la section "Examen de SEP0832" sur moodle :

- les seuls documents autorisés pour cette épreuve sont les CM et TD du module SEP0832, disponibles sur moodle ;
- cette épreuve est **individuelle** ; toute **communication** avec un tiers, en lien avec le sujet, est **formellement interdite** ;
- pour les QCM, pour chaque question, **réécrire (à la main et en entier)** la ou les réponses choisie(s), parmi les réponses proposées.

Je serai joinable via Teams de préférence par chat, en cas de nécessité, pour répondre à vos questions éventuelles.

On cherche à expliquer/prédire une variable statistique Y à l'aide de p variables statistiques $X_1, \ldots, X_p$. Notons $\mathcal{Y}$ le domaine de Y, et $\mathcal{X}$ le domaine du vecteur statistique $(X_1, \ldots, X_p)$.

(1) La variable statistique Y s'appelle

- la variable réponse
- la variable dépendante
- la variable explicative
- la variable indépendante
- la variable expliquée
- la variable endogène
- la variable exogène

(2) Les variables statistiques $X_1, \ldots, X_p$ s'appellent

- les variables dépendantes
- les variables indépendantes
- les variables explicatives
- les régresseurs
- les prédicteurs
- les variables exogènes
- les variables endogènes

(3) On parle de problème de régression si

- la variable Y est quantitative continue
- la variable Y est quantitative discrète avec Card($\mathcal{Y}$) grand
- la variable Y est quantitative discrète avec Card($\mathcal{Y}$) petit
- la variable Y est qualitative nominale
- la variable Y est qualitative avec nombre de modalités fini et petit

(4) On parle de problème de classification supervisée (ou discrimination supervisée) si
- la variable Y est quantitative continue
- la variable Y est quantitative discrète avec Card($\mathcal{Y}$) grand
- la variable Y est quantitative discrète avec Card($\mathcal{Y}$) petit
- la variable Y est qualitative nominale
- la variable Y est qualitative avec nombre de modalités fini et petit

(5) En régression linéaire, les variables $X_1, \ldots, X_p$, peuvent être
- numériques
- qualitatives
- qualitatives, si chacune des variables qualitatives est remplacée par les variables indicatrices (de ses modalités sauf une)
- quantitatives continues
- quantitatives discrètes

Supposons dorénavant que toutes les variables statistiques Y, $X_1, \ldots, X_p$ soient numériques. Notons $\mathbf{X} := (X_1, \ldots, X_p) \in \mathbb{R}^p$, le vecteur aléatoire prenant ses valeurs dans $\mathbb{R}^p$.

Considérons le modèle de régression linéaire multiple (RLM)

$$Y = w_0 + w_1 X_1 + \ldots + w_p X_p + \varepsilon. \tag{1}$$

On dispose de n ($n > p$) observations $(\mathbf{X}_1, Y_1) \in \mathbb{R}^{p+1}, \ldots, (\mathbf{X}_n, Y_n) \in \mathbb{R}^{p+1}$ i.i.d. de même loi que $(\mathbf{X}, Y) \in \mathbb{R}^{p+1}$.

On note

$$\mathbf{Y} := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbb{X} := \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix}, \ \boldsymbol{\varepsilon} := \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \ \boldsymbol{w} := \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}.$$

Notons $\widehat{\boldsymbol{w}}$ l'estimateur des moindres carrés du paramètre vectoriel $\boldsymbol{w}$, du modèle linéaire multiple (1) ci-dessus.

(6) L'estimateur $\widehat{\boldsymbol{w}}$ existe et est unique si
- les colonnes de $\mathbb{X}$ sont non corrélées (deux-à-deux)

- les colonnes de $\mathbb{X}$ sont linéairement indépendantes
- la matrice $\mathbb{X}$ est de plein rang
- la matrice $\mathbb{X}$ est de rang $(1+p)$
- la matrice $\mathbb{X}^\top \mathbb{X}$ est de plein rang
- la matrice $\mathbb{X}^\top \mathbb{X}$ est de rang p

(7) Lorsque le nombre de variables dans le modèle RLM est petit
- la variance est élevée
- la variance est faible
- le biais est élevé
- le biais est faible
- l'erreur empirique (d'ajustement) est faible
- l'erreur empirique (d'ajustement) est élevée

(8) Lorsque le nombre de variables dans le modèle de RLM est grand
- la variance est élevée
- la variance est faible
- le biais est élevé
- le biais est faible
- l'erreur empirique (d'ajustement) est faible
- l'erreur empirique (d'ajustement) est élevée

(9) On veut tester la validité du modèle de RLM complet (1) précédent.
  (a) Écrire l'hypothèse nulle $\mathcal{H}_0$ et l'hypothèse alternative $\mathcal{H}_1$ (en termes des paramètres du modèle) ;
  (b) Pour tester ces hypothèses, on utilise la statistique de Fisher suivante

$$F := \frac{R^2}{1-R^2}\frac{n-p-1}{p}$$

  qui suit une loi de Fisher, à p et $n-p-1$ degrés de liberté, sous l'hypothèse nulle.
  Écrire la forme explicite du R^2 ;
  Écrire la P-value de ce test.

(10) On veut tester si la variable $X_1 \in \mathbb{R}$ est significative dans le modèle de RLM précédent.
  (a) Écrire l'hypothèse nulle $\mathcal{H}_0$ et l'hypothèse alternative $\mathcal{H}_1$ (en termes des paramètres du modèle) ;
  (b) Pour réaliser ce test, on peut utiliser la statistique de Student

$$t := \frac{\widehat{w}_1}{\widehat{\sigma}_{\widehat{w}_1}}$$

  qui suit une loi de Student, à $n-p-1$ degrés de liberté, sous $\mathcal{H}_0$.
  Que représente le terme $\widehat{\sigma}_{\widehat{w}_1}$? Expliquer comment l'obtenir ;
  Écrire la P-value de ce test.

(c) Pour tester les hypothèses précédentes, on peut utiliser aussi la statistique de Fisher, notée F, (entre modèles emboîtés), qui suit une loi de Fisher à 1 et $n-p-1$ degrés de liberté, sous $\mathcal{H}_0$.
Écrire la statistique de test ;
Écrire la P-value correspondante.

(d) Parmi les deux tests statistiques (de Student ou Fisher) précédents, lequel choisiriez-vous ? Justifier.

(11) Supposons que $p \geq 3$. On se propose de tester si les variables explicatives X_1 et X_2 sont simultanément non significatives, dans le modèle complet.

(a) Écrire l'hypothèse nulle $\mathcal{H}_0$ et l'hypothèse alternative $\mathcal{H}_1$ (en termes des paramètres du modèle) ;

(b) Écrire la statistique de Fisher de ce test ;

(c) Donner sa loi sous l'hypothèse nulle $\mathcal{H}_0$;

(d) Écrire la P-value correspondante.

(12) En régression Lasso, si le paramètre de pénalisation λ augmente, alors

- la variance diminue
- la variance augmente
- le biais diminue
- le biais augmente
- le nombre de coefficients nulls augmente
- le nombre de coefficients nulls diminue

(13) Pour estimer l'erreur de prévision d'un modèle de RLM, on peut utiliser

- l'erreur empirique (d'ajustement)
- les méthodes de validation croisée

(14) Pour l'estimation de l'erreur de prévision d'un modèle (de RLM) par les méthodes de type validation-croisée

(a) citer le(s) avantage(s)/inconvénient(s) de la méthode apprentissage-validation ;

(b) citer le(s) avantage(s)/inconvénient(s) de la méthode leave-one-out cross-validation.