UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

Master 1 SEP
Examen de SEP0832
Méthodes d'échantillonnage
Durée 1h30
Session 2
Responsable : Amor KEZIOU
Juin 2019

Faculté des Sciences Reims

2018-2019

Aucun document n'est autorisé
Téléphones et calculatrices interdits

Nom : ..... **Prénoms** : .....

QCM : Cocher la ou les réponses correcte(s) parmi les choix proposés ;
Pour chacune des questions, cocher une mauvaise réponse conduit à une note zero ;
Le barème est de 1 point pour les questions à une seule réponse correcte, et de 0,5 point par réponse correcte pour les questions à plusieurs réponses correctes.

On cherche à expliquer/prédire une variable statistique Y par 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 $\mathrm{Card}(\mathcal{Y})$ grand

☐ la variable Y est quantitative discrète avec $\mathrm{Card}(\mathcal{Y})$ 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 $\mathbf{X} := (X_1, \ldots, X_p)$.

Considérons le modèle de régression linéaire multiple (RLM)

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

On dispose de n observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ i.i.d. de même loi que $(\mathbf{X}, Y)$.

On note

$$\mathbf{Y} := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbb{X} := \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix}, \boldsymbol{\varepsilon} := \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \ \boldsymbol{\beta} := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Notons $\widehat{\boldsymbol{\beta}}$ l'estimateur des moindres carrés du modèle linéaire multiple ci-dessus.

(6) L'estimateur $\widehat{\boldsymbol{\beta}}$ 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 global du modèle de RLM précédent.

a) Écrire l'hypothèse nulle et l'hypothèse alternative (en termes des paramètres du modèle) :

...

...

b) Pour réaliser ce test, 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 P-value de ce test :

...

...

(10) On veut tester si la variable X_p est significative dans le modèle de RLM précédent.

a) Écrire l'hypothèse nulle et l'hypothèse alternative (en termes des paramètres du modèle) :

...

...

b) Pour réaliser ce test, on peut utiliser la statistique de Student suivante

$$t := \frac{\widehat{\beta}_p}{\widehat{\sigma}_{\widehat{\beta}_p}}$$

qui, sous l'hypothèse nulle, suit $\mathcal{T}(n-p-1)$, une loi de Student à $n-p-1$ degrés de liberté. Écrire la P-value de ce test :

...

...

c) Pour réaliser le test précédent, on peut utiliser également la statistique de Fisher (entre modèles emboîtés) suivante

$$F := \frac{\|\widehat{\mathbf{Y}}_0 - \widehat{\mathbf{Y}}\|}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|/(n-p-1)}$$

qui, sous l'hypothèse nulle, suit $\mathcal{F}(1, n-p-1)$, une loi de Fisher à 1 et $n-p-1$ degrés de liberté. Écrire la P-value de ce test :

...

...

d) Parmi les deux tests statistiques précédents (de Student et Fisher), 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 significatives simultanément.

a) Écrire l'hypothèse nulle et l'hypothèse alternative (en termes des paramètres du modèle) :

...

...

b) Pour réaliser ce test, on peut utiliser la statistique de Fisher suivante

$$F := \frac{\|\widehat{\mathbf{Y}}_0 - \widehat{\mathbf{Y}}\|/2}{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|/(n-p-1)}.$$

Cette statistique suit, sous l'hypothèse nulle, $\mathcal{F}(2, n-p-1)$, une loi de Fisher à 2 et $n-p-1$ degrés de liberté. Écrire la P-value de ce test.

...

...

(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) En régression Lasso, si le paramètre de pénalisation λ diminue, 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

(14) En régression Ridge, 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

(15) En régression Ridge, si le paramètre de pénalisation λ diminue, 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

(16) Pour estimer l'erreur de prévision d'un modèle de RLM, on peut utiliser
- ☐ l'erreur empirique (d'ajustement)
- ☐ le bootstrap
- ☐ les méthodes de validation croisée

(17) Citer les inconvénients de la méthode "apprentissage-validation" (en sélection de modèles)

...

...

...

(18) Citer les inconvénients de la méthode "leave-one-out" (en sélection de modèles)

...

...

...

(19) Citer les avantages de la méthode "K-fold cross-validation" ("$K = 10$", en sélection de modèles)

...

...

...