S3 4 Wrap Up Régressions

Table of contents

1 Régressions linéaires multiples

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- Y : variable **dépendante** (quantitative continue) : par exemple la pression artérielle
- X1, X2, …, Xk : variables **indépendantes** :
  - peuvent être quantitatives continues (âge, IMC, etc.)
    - dans le cas de l'âge : ça s'exprime "de combien augmente la valeur de Y pour une augmentation d'une unité d'âge (1 an)"
  - peuvent être binaires
  - pour les variables catégorielles ordonnées : elles sont transformées en variables quantitatives !
- ε : terme d'erreur

Et donc : β1 correspond à la différence du facteur X1 (par ex âge) entre sujets de facteur X2 constant (par ex sexe)

On peut tester :

- H0 : β1 = 0 (pas d'association entre X1 et Y)
- H1 : β1 ≠ 0 (association entre X1 et Y)

Interprétation d'un coefficient βi de régression :

- Si Xi est une variable quantitative continue : βi correspond à la variation moyenne de Y associée à une augmentation d'une unité de Xi, toutes les autres variables Xj (j ≠ i) étant maintenues constantes.
- Si Xi est une variable binaire (0/1) : βi correspond à la différence moyenne de Y entre les sujets avec Xi = 1 et les sujets avec Xi = 0, toutes les autres variables Xj (j ≠ i) étant maintenues constantes.

1.A Conditions de validité :

Du test β1 = 0 :

- ε soit un bruit d**e variance constante** et **suivant une loi normale**
- donc idem qu'un test t !!
- pour vérifier : histogramme des résidus !

Définition des résidus :

$$\varepsilon_i = Y_i - \hat{Y}_i$$

- pour chaque individu i, le modèle prédit une valeur Ŷi (ce que le modèle estime pour Yi)
- la valeur Yi correspond à la valeur réelle observée pour l'individu i
- Le résidu pour l'individu i est : ε = Y – Ŷ
- Le résidu mesure **l'erreur du modèle** pour l'individu i.
- Si proche de 0 : le modèle prédit bien la valeur observée
- Si grand (positif ou négatif) : le modèle prédit mal la valeur observée
- Pour vérifier les conditions de validité :
  - tracer les résidus en fonction des valeurs prédites Ŷi
  - vérifier la normalité des résidus (histogramme, Q-Q plot)

hétéroscédasticité : variance des résidus non constante

1.B Corrélatons entre variables explicatives

$$Y = \beta_0 + \beta_1(X_1X_2) + \beta_3X_3 + \cdots + \beta_kX_k + \varepsilon$$

- Par ex β1 : âge, β2 : IMC.
- Effet **multiplicatif** (par interaction : encore + que additif)
- Recherche une **synergie** entre les variables explicatives

Synergie = non linéarité, donc c'est rare de mettre en évidence une non-linéarité en médecine.

Mais en vrai, les FDR multipliés se potentialisent !

ATTENTION : si on met des interactions dans le modèle, on ne peut plus interpréter les coefficients βi de façon isolée en dehors du facteur β1 !

Le mieux :

- D'abord faire un modèle sans interaction
- Puis faire un modèle avec interaction

Mais pour les esthètes de la statistiques : le modèle avec interaction ne peut pas vraiment être utilisé car les résidus sont multipliés, donc c'est pas des vrais résidus. Bruno dit que ce n'est pas grave.

1.B.1 Exemple d'interprétation avec interaction

$Y = \beta_0 + \beta_1 \text{ retraite} + \beta_2 \text{ surpoids} + \beta_3(\text{âge} \times \text{IMC}) + \varepsilon$

	OR et p value	Interprétation
β1		**Ininterprétable**
β2		**Ininterprétable**
β3		OUI

Mais si c'est codé en -1 / 1 (pour FALSE - TRUE) : Alors **β1 et β2 pourraient être interprétable** !

- **β1** est un effet de l'âge dans une population où il y aurait autant de surpoids que de pas de surpoids
- **β2** est un effet du surpoids dans une population où il y aurait autant de vieux que de jeunes
- **β3** ne change pas que ce soit en -1 / 1 ou 0 / 1 : reste l'effet d'interaction entre âge et IMC
- NB : si variable quantitative : ne pas le faire

Pour ça dans R, il faut changer le contraste avec la fonction `contr.sum()` Documentation R

1.B.2 Si X1 était catégorielle à plus de 2 modalités ?

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$

- Y : Tension artérielle
- X1 : surpoids
- X2 : profession (catégorielle à 4 modalités : cadre, intermédiaire, employé, ouvrier) : variable pas quantitative, pas ordonnée, pas binaire.

Il faut recoder X2 en variables binaires (variables indicatrices) (besoin d'en faire que 3 variables binaires pour 4 modalités) :

Profession	X2_cadre	X2_intermédiaire	X2_employé
Cadre	1	0	0
Intermédiaire	0	1	0
Employé	0	0	1

Profession	X2_cadre	X2_intermédiaire	X2_employé
Ouvrier	0	0	0

Donc le modèle devient :

$$Y = \beta_0 + \beta_1 \text{ surpoids} + \beta_2\, X2_{\text{cadre}} + \beta_3\, X2_{\text{intermédiaire}} + \beta_4\, X2_{\text{employé}} + \varepsilon$$

Et on peut même résumer : β2 est la différence moyenne de tension artérielle entre les cadres et les ouvriers (catégorie de référence), toutes les autres variables Xj (j ≠ 2) étant maintenues constantes.

Pour choisir la profession de référence : (fonction `relevel()` dans R)

- soit la catégorie la plus fréquente
- soit la catégorie la plus "neutre"
- mais le logiciel en choisit une par défaut (souvent la première par ordre alphabétique)

Pour éviter d'avoir une catégorie de référence :

- Exemple : comparaison inter-hôpitaux
- Utiliser des **contrastes**

Par exemple comparaison Lariboisière, Pompidou, Bichat, Beaujon

tableau :

Hôpital	X2_Lariboisière	X2_Pompidou	X2_Bichat
Lariboisière	1	0	0
Pompidou	0	1	0
Bichat	0	0	1
Beaujon	0	0	0

On crée des variables indicatrices mais on les code différemment (avec des -1 et des 1) :

Hôpital	X2_Lariboisière	X2_Pompidou	X2_Bichat	X2_Beaujon
Lariboisière	-1	0	0	1
Pompidou	0	-1	0	1
Bichat	0	0	-1	1
Beaujon	1	1	1	1

Dans ce cas, X2 est la différence de TAmoyenne en comparaison à l'effet moyen de tous les hôpitaux.

Si nécessaire : vidéo à ≈ 30 minutes

1.C Réponse aux questions

Modèle : Y = a0 + a1(X1*X2)

Dans ce cas : interprétation difficile !!

Le mieux est de faire $Y = a0 + a1X1 + a2X2 + a3(X1*X2)$

2 Régression logistique

Modélisation d'une variable dépendante binaire (0/1).

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

avec :

- p : probabilité que maladie = 1 (événement)
- (1-p) : probabilité que maladie = 0 (non-événement)
- log(p/(1-p)) : logit de p

Sinon idem que pour la régression linéaire multiple mais pas de bruit ε (car variable dépendante binaire et on modélise la probabilité p).

Ici :

- Y = maladie (0/1) pour HTA oui / non
- β_1 : effet du surpoids sur l'HTA ou non
  - et $exp(\beta_1)$ = OR ajusté sur β_2 de β_1 par rapport à l'HTA.
  - c'est à dire : si le surpoids était rare (donc OR ≈ RR) et $exp(\beta_1) = 2$, la probabilité d'avoir une HTA si l'on est en surpoids = 2.

2.A Interactions

Exactement de la même manière qu'en régression linéaire ! et possible de recoder en -1 / 1 pour pas avoir de variable de référence.

2.B Conditions de validité de la régression logistique

- Taille de l'échantillon suffisante (10 événements par variable explicative au minimum)

2.C Autres trucs :

- Test de calibration / Hosmer-Lemeshow : accepter l'hypothèse nulle alors qu'on a pas la puissance !! Donc en théorie on ne peut pas vraiment faire ce test

3 Références

Vidéo Falissard :

- Blog larmarange
- Webin-R