Analyse multivariée

Filière Geo Data Science : UE2 Analyse de Données

Juste Raimbault[1]

2024-2025

[1]LaSTIG, IGN-ENSG-UGE

ENSG Géomatique
ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

Contenu du cours

Régression multiple

Séries temporelles

Régression multiple

Modèle linéaire multiple

Pour une variable indépendante Y et N variables explicatives X_i

$$Y = \beta_0 + \sum_{i=1}^{N} \beta_i \cdot X_i + \varepsilon$$

avec ε une erreur aléatoire indépendante et distribuée identiquement entre les différentes observations

Minimisation de la somme des erreurs au carré :

$$S(\beta) = {}^t(\vec{y} - \vec{\beta} \cdot \mathsf{X})(\vec{y} - \vec{\beta} \cdot \mathsf{X})$$

qui admet un minimum global en

$$\hat{\beta} = ({}^t\mathsf{X} \cdot \mathsf{X})^{-1}({}^t\mathsf{X}) \cdot \vec{y}$$

- capture des relations linéaires → toujours visualiser les données pour ne pas manquer des relations non-linéaires flagrantes
- pas de causalité
- hypothèse *IID* → visualiser les résidus
- qualité de la régression : R2 ajusté
- valeur des coefficients : force de l'effet de la variable
- p-value des coefficients : hypothèse nulle "Coefficient égal à 0" (test de Student)

Sélection des variables

Quel modèle choisir ? Problème du sur-ajustement

Critère d'Information d'Akaike pour estimer la quantité d'information extraite par le modèle, avec k paramètres estimés et L vraisemblance maximale

$$AIC = 2k - 2\ln L$$

$\rightarrow$ critère à minimiser, comparable uniquement pour des modèles sur les même jeux de données

$\rightarrow$ application à la sélection des variables explicatives dans la régression multiple

Régression pas à pas

Pour faire de la sélection selon un critère si tester l'ensemble des modèles n'est pas possible :

- *Forward selection* : en partant du modèle le plus simple, on ajoute la variable qui améliore le plus à chaque étape
- *Backward elimination* : en partant du modèle complet, on élimine la variable la moins importante à chaque étape
- *Stepwise regression* : combinaison des deux (en ascendant ou descendant), remise en cause des variables déjà incluses à chaque étape dont le rôle peut changer avec des contrôles supplémentaires

→ implémentation en R avec la fonction `step`

Termes d'interaction et polynômes

Modèle avec interactions au premier ordre :

$$Y = \beta_0 + \sum \beta_i X_i + \sum_{k \neq l} \beta_{kl} X_k X_l + \varepsilon$$

$\rightarrow$ généralisation avec n'importe quel polynôme en les variables X_i

$\rightarrow$ souvent peu significatifs (méthodes spécifiques pour tester la puissance statistique), des méthodes non-paramétriques peuvent être préférées

Effets fixes

$\rightarrow$ ajout de termes constants en fonction de groupes

$\rightarrow$ permet de contrôler avec une variable discrete (individu dans des données panel, jour ou année, pays ou ville, . . .)

$\rightarrow$ problème : très haute dimension rapidement selon le nombre de modalités

$\rightarrow$ vu plus en détails avec les régression multi-niveaux en analyse spatiale

Modèles linéaires généralisés (GLM)

Prise en compte d'une non-linéarité avec une fonction lien g entre la composante linéaire et la moyenne de la réponse, avec une variance non homogène :

$$\mathbb{E}(Y|\mathsf{X} = g^{-1}(\mathsf{X}\vec{\beta})$$

$\rightarrow$ estimation souvent par maximum de vraisemblance (pas d'expression exacte) : plus coûteux en calcul

$\rightarrow$ modèles de Poisson (comptage), logistique (résultat binaire), probit (choix discrets), ...

Exemple : modèle Logit

Fonction de lien avec une courbe logistique, pour estimer une probabilité p de succès/échec :

$$\ln \frac{p}{1-p} = \beta_0 + \sum \beta_i X_i$$

$\rightarrow$ classifieur binaire simple

$\rightarrow$ implémentation en R dans la fonction `glm`

Problème du nombre de variables explicatives en très haute dimension, avec des données *sparse*

$\rightarrow$ pénalisation de l'erreur avec un terme de norme du vecteur des coefficients

$\rightarrow$ Ridge regression : pénalisation avec la norme L2

$\rightarrow$ Lasso regression : pénalisation avec la norme L1, i.e. ajout d'un terme

$$\lambda \sum |\beta_i|$$

dans l'erreur à minimiser

Séries temporelles

Séries temporelles

Avec des données temporelles, certaines propriétés ne sont plus vérifiées (indépendance des erreurs, uniformité de la variance) et ce qu'on cherche à modéliser peut être différent

$\rightarrow$ méthodes statistiques spécifiques au séries temporelles

$\rightarrow$ modèles d'autorégression : AR, ARCH, ARMA, ARIMA

$\rightarrow$ méthodes pour estimer des causalités

Régression linéaire sur deux séries stationnaires, en incluant les variables retardées :

$$Y_t = y_0 + \sum \beta_\tau^{(Y)} Y_{t-\tau} + \sum \theta_\tau^{(Y)} X_{t-\tau} + \varepsilon_t^{(Y)}$$

$$X_t = x_0 + \sum \beta_\tau^{(X)} X_{t-\tau} + \sum \theta_\tau^{(X)} X_{t-\tau} + \varepsilon_t^{(X)}$$

$\rightarrow$ test statistique sur les coefficients θ, causalité de Granger ("faible") si X aide à prédire Y ou réciproquement

$\rightarrow$ nombreuses autres méthodes de causalité : entropie de transfert, variables instrumentales, diff-in-diff, contrôle synthétique, ...