Université Paris 1 Panthéon-Sorbonne

UFR02 Économie

Sorbonne Data Analytics

PROJET D'ÉCONOMÉTRIE APPLIQUÉE

Analyse des Prix Immobiliers

Du modèle linéaire aux méthodes de régularisation

Durée estimée : 3-4 semaines

Travail : En groupes de 2-3 étudiants

Remise : Rapport écrit + présentation (PPTX)

Année universitaire 2025-2026

Table des matières

Contexte

Vous disposez d'un ensemble de données sur 150 maisons vendues dans une région entre 2015 et 2023. Votre objectif est d'analyser les facteurs qui influencent le prix de ces maisons en utilisant l'ensemble des outils économétriques vus en cours.

Dataset : `donnees_immobilieres.xlsx`

Variables disponibles :

Variable	Description
`ID`	Identifiant unique de la maison
`Surface_m2`	Surface habitable en mètres carrés
`Chambres`	Nombre de chambres
`Annee_construction`	Année de construction
`Distance_centre_km`	Distance au centre-ville en kilomètres
`Etage`	Étage (0 = rez-de-chaussée)
`Ascenseur`	Présence d'un ascenseur (1 = oui, 0 = non)
`Annee_vente`	Année de la vente (2015-2023)
`Qualite_ecole`	Score de qualité des écoles du quartier (1-10)
`Revenu_median_quartier`	Revenu médian du quartier (en milliers €)
`Prix_milliers_euros`	Prix de vente en milliers d'euros *(Variable dépendante)*

1 Statistiques Descriptives et Analyse Préliminaire

1.1 Statistiques descriptives

Calculez et présentez les statistiques descriptives pour chaque variable :

- Moyenne ($\bar{X}$), médiane, écart-type (s_X)
- Minimum, maximum, quartiles
- Asymétrie (*skewness*) et aplatissement (*kurtosis*) pour le prix
- Présentez un tableau récapitulatif

À faire

Créez des histogrammes et des boîtes à moustaches pour visualiser les distributions. Identifiez les variables qui pourraient nécessiter une transformation logarithmique.

1.2 Analyse de corrélation

- Calculez la **matrice de corrélation** entre toutes les variables continues
- Créez un **graphique de corrélation** (heatmap)
- Identifiez les paires de variables fortement corrélées entre elles (risque de multicolinéarité)

Question

Quelle variable semble avoir l'impact le plus fort sur le prix selon la corrélation ? Attention : corrélation $\neq$ causalité !

2 Le Modèle Linéaire : Estimation et Interprétation

2.1 Modèle de régression linéaire simple

Première étape : Régressez le prix sur la surface uniquement.

$$\text{Prix}_i = \beta_0 + \beta_1 \times \text{Surface}_i + u_i \tag{1}$$

- Estimez les coefficients par **MCO**
- Présentez les résultats avec :
  - Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$
  - L'écart-type de chaque coefficient ($\hat{\sigma}_{\hat{\beta}_j}$)
  - La statistique t et la p-valeur
  - Le R^2 et le R^2 ajusté

Question

Interprétation : Que signifie le coefficient $\hat{\beta}_1$? Si la surface augmente de 1 m², de combien le prix augmente-t-il en moyenne ?

2.2 Modèle de régression linéaire multiple

Spécification du modèle

$$\begin{aligned}\text{Prix}_i = \beta_0 + \beta_1 \times \text{Surface}_i + \beta_2 \times \text{Chambres}_i + \beta_3 \times \text{Annee\_construction}_i \\ + \beta_4 \times \text{Distance\_centre}_i + \beta_5 \times \text{Etage}_i + \beta_6 \times \text{Ascenseur}_i + u_i\end{aligned} \tag{2}$$

Estimez le modèle et présentez les résultats

1. Tous les coefficients sont-ils significatifs ?
2. Quel est l'impact marginal de chaque variable sur le prix ?
3. Pour la variable `Ascenseur` : comment interpréter le coefficient ?
4. Comment interprétez-vous la différence entre R^2 et $\bar{R}^2$?

2.3 Transformation logarithmique (5 points)

Modélisez en semi-log et en log-log.

1. Comparez les trois modèles.
2. Quel modèle semble le plus approprié et pourquoi ?

3 Diagnostics du Modèle

3.1 Multicolinéarité

- Calculez les **VIF** (Variance Inflation Factor) pour chaque variable

Question

Y a-t-il des variables avec un VIF élevé ? Faut-il en supprimer certaines ? Definir le biais de variable omise.

4 Tests et Inférence

1. Testez l'hypothèse que la distance au centre a un effet négatif sur le prix. Quelle est la p-value ?
2. Testez l'hypothèse que tous les coefficients (sauf constante) soient nuls. Testez si l'ajout des variables : `Qualite_ecole` et `Revenu_median_quartier` améliore significativement le modèle
3. Pourquoi ne peut-on pas simplement utiliser plusieurs tests T pour tester plusieurs restrictions simultanément ?

4.1 Stabilité structurelle

Testez si le COVID a un effet sur le marché immobilier en utilisant la méthode de votre choix.

À faire

Si vous trouvez une rupture structurelle, discutez des implications pour votre analyse. Faut-il estimer des modèles séparés ?

5 Hétéroscédasticité et Autocorrélation

1. Observez graphiquement si les résidus suivent un pattern.
2. Testez l'hétéroscédasticité et corrigez-la.
3. Comparez les MCO standard, les MCO avec écarts-types robustes, et WLS.

5.1 Test d'autocorrélation

Testez l'autocorrélation

À faire

Si vous détectez à la fois hétéroscédasticité et autocorrélation, utilisez les **écarts-types de Newey-West** qui sont robustes aux deux problèmes.

6 Endogénéité et Variables Instrumentales

6.1 Sources d'endogénéité

Quelles sont les sources possibles d'endogénéité dans notre contexte ?

Question

Discutez : la variable `Qualite_ecole` est-elle potentiellement endogène ? Pourquoi ?

6.2 Estimation par Variables Instrumentales

Proposition d'instrument

Introduisez la variable `Distance_universite` (distance à l'université la plus proche).

1. Argumentez pourquoi cette variable pourrait être un bon instrument pour `Qualite_ecole`.
2. Construisez une estimation en deux étapes (2SLS)
3. Testez la validité des instruments avec la méthode de votre choix.
4. Comparez les coefficients MCO et IV. Y a-t-il des différences importantes ?

7 Régularisation

À faire

Important : Avant d'appliquer Ridge ou Lasso, standardisez toutes les variables (moyenne 0, écart-type 1) car la pénalité dépend de l'échelle des coefficients.

1. Estimez un modèle Ridge avec différentes valeurs de λ. Analysez et commentez l'évolution des coefficients.
2. Estimez un modèle Lasso pour différentes valeurs de λ. Analysez et commentez la manière dont les coefficients se modifient en fonction de λ.
3. Choissisez la valeur du paramètre λ. Pour cela, utilisez la validation croisée 10-fold pour choisir λ optimal.
4. Comparez les résultats de trois modèles sur votre jeu de données. Divisez en train et test (80% - 20%) et comparez les erreurs de prédiction (RMSE) sur l'échantillon de test.

Question

Discussion : Pourquoi les écarts-types et tests classiques ne sont-ils pas valides après Lasso ?

8 Prévisions

8.1 Prédiction ponctuelle et intervalle de confiance

Prédisez le prix d'une maison avec les caractéristiques suivantes :

- Surface : 120 m²
- Chambres : 3
- Année construction : 2015
- Distance centre : 5 km
- Étage : 1
- Ascenseur : Oui
- Année vente : 2023
- Qualité école : 7
- Revenu médian quartier : 65 000 €
- Distance université : 4 km

1. Calculez la prédiction ponctuelle avec votre meilleur modèle
2. Calculez l'intervalle de confiance à 95%
3. Cette prédiction est-elle fiable ? Discutez.

Rapport à Remettre

Structure obligatoire

1. **Page de titre** (nom du groupe, date, titre)
2. **Résumé exécutif** (1 page max)
   - Principaux résultats
   - Recommandations
3. **Introduction** (1-2 pages)
   - Contexte et problématique
   - Structure du rapport
4. **Partie 1 : Analyse descriptive et modèle de base** (5-7 pages)
   - Statistiques descriptives
   - Modèle linéaire simple et multiple
   - Tests de significativité
5. **Partie 2 : Diagnostics et corrections** (7-10 pages)
   - Multicolinéarité et observations influentes
   - Tests d'hétéroscédasticité et corrections
   - Test de Chow
6. **Partie 3 : Endogénéité** (5-7 pages)
   - Discussion des sources potentielles
   - Estimation IV (si applicable)
   - Tests de validité
7. **Partie 4 : Méthodes de régularisation** (5-7 pages)
   - Ridge et Lasso
   - Comparaison des performances prédictives
8. **Conclusion et recommandations** (2-3 pages)
   - Synthèse des résultats
   - Limites de l'analyse
   - Recommandations pour la pratique
9. **Annexes**
   - Tableaux complets
   - Code (Python)
   - Graphiques supplémentaires