

Méthodes Quantitatives

Séance 10 : Introduction à la régression linéaire

Analyse des relations entre variables

Sciences Po

2024-2025

Plan de la séance

- 1 Introduction : pourquoi la régression ?
- 2 Histoire de la régression
- 3 Principes de la régression linéaire simple
- 4 Interprétation des résultats
- 5 Extensions de la régression linéaire
- 6 Pièges et précautions
- 7 Variables instrumentales et équations
- 8 En pratique
- 9 Conclusion

Introduction : pourquoi la régression ?

De la corrélation à la régression

- **Limites de la corrélation :**

- Mesure uniquement la force du lien
- Ne permet pas de prédictions
- Ne tient pas compte de la causalité

- **Avantages de la régression :**

- Modélise la relation entre variables
- Permet des prédictions
- Peut intégrer plusieurs variables

Questions typiques en sciences sociales

- **En économie :**

- Impact du niveau d'éducation sur le salaire
- Effet du taux d'intérêt sur l'investissement

- **En sociologie :**

- Influence du milieu social sur la réussite scolaire
- Déterminants des pratiques culturelles

- **En science politique :**

- Facteurs du vote
- Déterminants de la participation électorale

Histoire de la régression

Francis Galton et les origines de la régression

- Contexte : Angleterre, fin du XIXe siècle
- Étude des caractéristiques héréditaires
- Observation : "régression vers la moyenne"
- Premier usage du terme "régression"
 - Taille des parents vs enfants
 - Les enfants "régressent" vers la moyenne

Développement historique

- **Karl Pearson** (élève de Galton)
 - Formalisation mathématique
 - Introduction du coefficient de corrélation
 - Développement des outils statistiques
- **Ronald Fisher**
 - Méthode des moindres carrés
 - Tests de signification
 - Applications en agronomie
- **Évolutions récentes**
 - Développement du calcul informatique
 - Nouvelles méthodes d'estimation
 - Applications en machine learning

Un héritage complexe

- **Contexte historique**

- Darwinisme social
- Eugénisme
- Déterminisme biologique

- **Leçons pour aujourd'hui**

- Impact social des méthodes quantitatives
- Responsabilité des chercheurs
- Importance de l'éthique

Principes de la régression linéaire simple

Le nuage de points

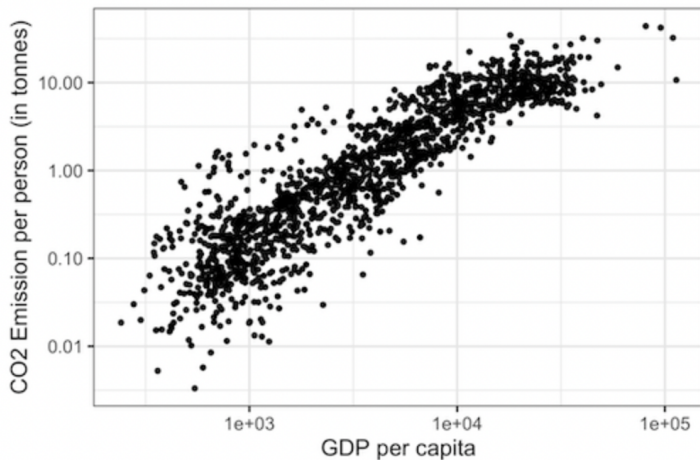
- **Définition**

- Représentation graphique de deux variables
- Chaque point = une observation
- Axes = variables X et Y

- **Interprétation**

- Forme du nuage
- Direction de la relation
- Dispersion des points

CO2 emission per person vs GDP per capita



La droite de régression

Équation de la droite

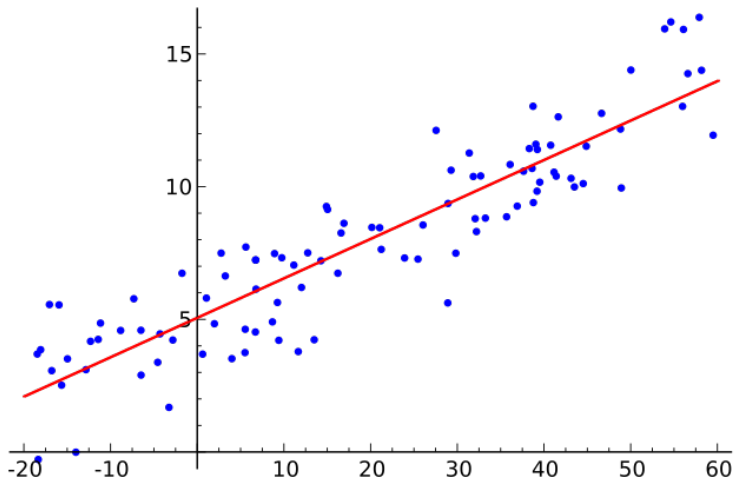
$$Y = aX + b + \epsilon$$

- **Paramètres :**

- a : pente (coefficient directeur)
- b : ordonnée à l'origine
- ϵ : terme d'erreur

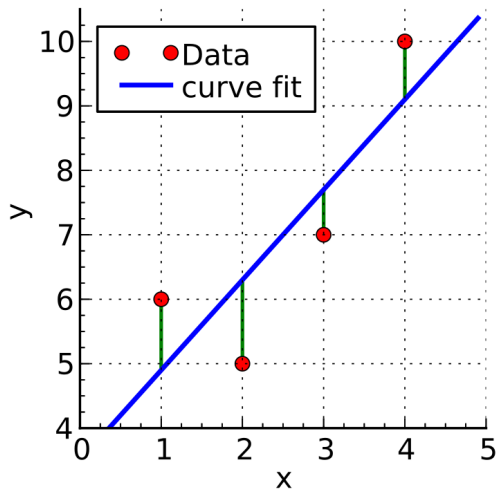
- **Interprétation :**

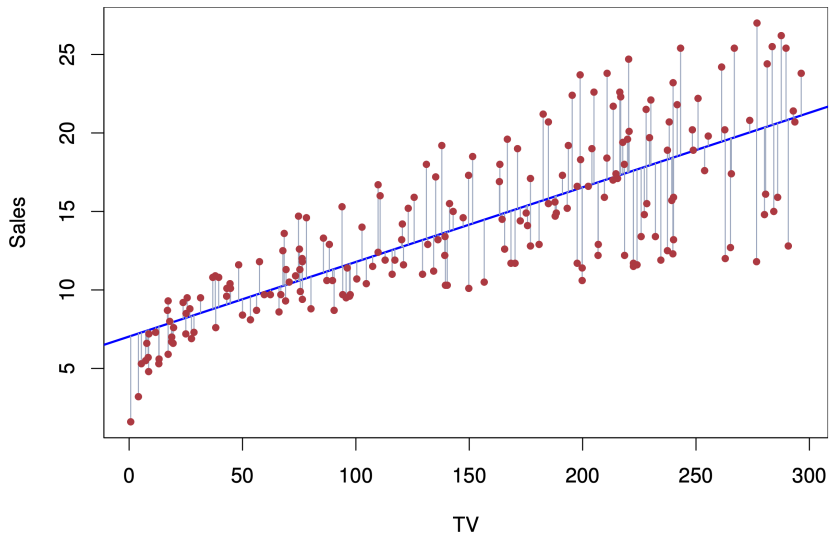
- a : variation de Y pour une unité de X
- b : valeur de Y quand $X = 0$



Principe des moindres carrés

- **Objectif** : trouver la "meilleure" droite
- **Méthode** :
 - Calculer les écarts verticaux
 - Minimiser la somme des carrés
- **Formellement** :
 - $\min \sum (Y_i - (aX_i + b))^2$





Hypothèses du modèle

- **Linéarité**

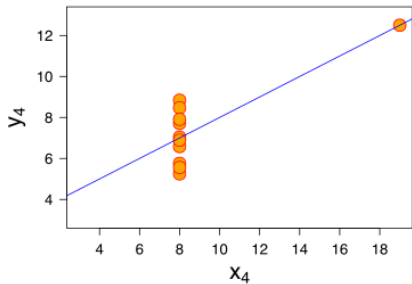
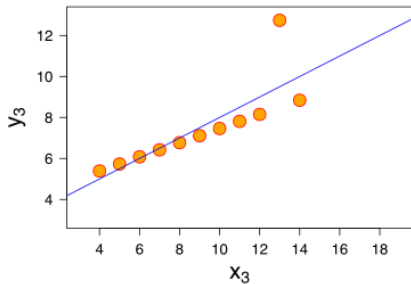
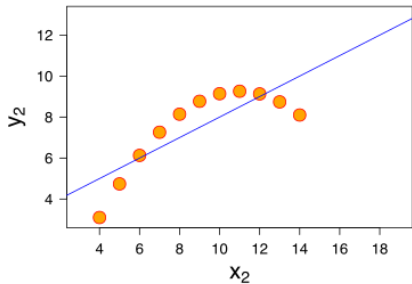
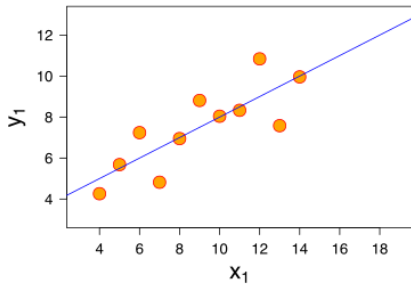
- Relation linéaire entre X et Y
- Vérification graphique importante

- **Indépendance des erreurs**

- Résidus non corrélés
- Important pour les séries temporelles

- **Homoscédasticité**

- Variance constante des résidus
- Vérification graphique



Interprétation des résultats

Interprétation des coefficients

- a : coefficient directeur de la droite de régression
 - Représente la variation moyenne de Y pour une variation de X de 1 unité
- b : ordonnée à l'origine
 - Valeur moyenne de Y lorsque $X = 0$

Le coefficient de détermination (R^2)

- **Définition :**

- Part de variance expliquée
- Varie de 0 à 1 (ou 0% à 100%)

- **Interprétation :**

- $R^2 = 0.25$: 25% de la variance expliquée
- Attention aux conclusions hâtives
- Dépend du domaine d'étude

Tests de significativité

- **Test sur les coefficients**

- $H_0 : \text{coefficient} = 0$
- P-valeur et seuils critiques

- **Notation conventionnelle**

- * : $p < 0.05$
- ** : $p < 0.01$
- *** : $p < 0.001$

Extensions de la régression linéaire

La régression multiple

- **Principe**

- Plusieurs variables explicatives
- $Y = a_1X_1 + a_2X_2 + \dots + b$

- **Avantages**

- Plus réaliste
- Contrôle des effets
- Meilleure explication

- **Difficultés**

- Interprétation plus complexe
- Problème de colinéarité
- Choix des variables

La colinéarité

- **Définition**

- Variables explicatives corrélées entre elles
- Redondance de l'information

- **Problèmes**

- Instabilité des coefficients
- Erreurs-types élevées
- Difficulté d'interprétation

La régression logistique

- **Cas particulier**

- Variable Y binaire (0/1)
- Ex : vote, emploi, diplôme

- **Principe**

- Modélise la probabilité
- Utilise la fonction logistique
- Coefficients = odds ratios, i.e. multiplication de la probabilité lorsque la variable augmente d'une unité.

La modalité de référence

- **Pour les variables catégorielles**

- Une modalité sert de référence
- Coefficients = comparaison avec référence

- **Choix important**

- Impact sur l'interprétation
- Souvent : modalité la plus fréquente
- Ou : modalité "naturelle" (ex : pas de diplôme)

Pièges et précautions

Application : Exemple du genre et du salaire

- **Question** : Les femmes sont-elles discriminées ?
- **Variables à considérer** :
 - Genre
 - Niveau d'éducation
 - Expérience professionnelle
 - Secteur d'activité
 - Temps de travail
- **Problèmes potentiels**
 - Colinéarité entre secteur et genre
 - Variables omises importantes
 - Biais de sélection

Variables instrumentales et équations

Corrélation n'est pas causalité

- **Exemples classiques**

- Ventes de glaces et noyades

- **Explications possibles**

- Variable cachée
- Causalité inverse
- Coïncidence

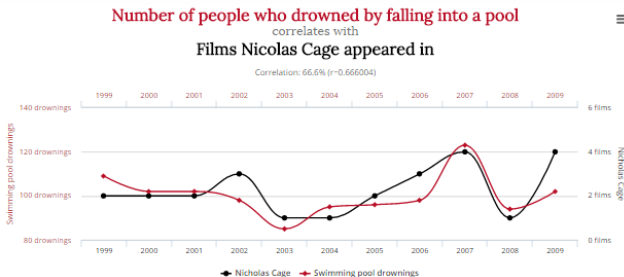
Corrélation n'est pas causalité

Exemples classiques

- Ventes de glaces et noyades

Explications possibles

- Variable cachée
- Causalité inverse
- Coïncidence



Variables Instrumentales : Une Introduction Simple

• Le problème

- Parfois, difficile de savoir si X cause vraiment Y
- Exemple : L'éducation augmente-t-elle le salaire ?
- Ou les personnes plus riches étudient-elles plus longtemps ?

• La solution : trouver un "instrument"

- Une variable Z qui influence X mais pas directement Y
- Exemple célèbre : distance à l'université
 - Influence les années d'études
 - Mais n'influence pas directement le salaire

• L'idée

- Utiliser seulement la partie de X influencée par Z
- Permet de mieux comprendre la vraie relation causale

En pratique

Manipulation dans Google Sheets

① Création du nuage de points

- Sélectionner les données
- Insertion > Graphique
- Type : Nuage de points

② Ajout de la droite de régression

- Personnaliser > Série
- Ajouter une courbe de tendance
- Afficher l'équation

La fonction LINEST()

- **Syntaxe :**

- =LINEST(y_known, x_known)
- Renvoie les coefficients de régression

- **Résultats :**

- Pente (a)
- Ordonnée à l'origine (b)
- R^2 et statistiques

L'article de Herpin (2003)

Question de recherche

Impact de la taille sur le revenu et la vie professionnelle

- **Données**

- Enquête EPCV 2001
- Échantillon représentatif
- Variables multiples

- **Méthodes**

- Régression linéaire multiple
- Contrôles sociodémographiques
- Tests de robustesse

Résultats principaux

- **Effet positif de la taille**

- +2.6% de revenu par cm supplémentaire
- Plus fort pour les hommes
- Persiste après contrôles

- **Mécanismes suggérés**

- Discrimination directe
- Effet sur la confiance en soi
- Impact sur les interactions sociales

Conclusion

Points clés à retenir

- **La régression comme outil**

- Méthode puissante mais avec limites
- Importance des hypothèses
- Nécessité d'interprétation prudente

- **En pratique**

- Toujours commencer par l'exploration
- Vérifier les hypothèses
- Interpréter avec le contexte

- **Pour aller plus loin**

- Autres types de régression
- Méthodes plus avancées
- Logiciels spécialisés