# Modèle de régression linéaire

*support 2012-2013*

*Objectif de la session:* Prise de contact avec le logiciel R. Régression linéaire multiple.

**Instructions R utilisées dans le TP:** read.table(); nrow(); ncol(); dim(); summary(); attributes(); mean(); sum(); length(); plot(); abline(); cor(); lm(); predict.lm(); solve(). L'opérateur $\% * \%$ effectue un produit matriciel. *Rappel:* `help(fonc)` pour obtenir de l'aide sur la fonction nommée "fonc".

## Méthode des moindres carrés Ordinaires -MCO

### Application

Le fichier "immo.txt" décrit un ensemble de transactions immobilières sur Paris. Les valeurs des variables sont sauvegardées par colonnes: col 1:"Surface du bien en $m^2$"; col 2: "valeur initiale d'achat "; col 3:"Montant de la transaction en K-euros";

- Charger le fichier dans l'environnement R et sauvegarder les données dans une structure dataframe nommée `tab` à l'aide de l'instruction `read.table()` en vous assurant que la première ligne du fichier de données définit les étiquettes du dataframe (tableau de données récupéré) Taper les instructions: `names(tab)`, `tab[,1]`, `tab$surface`, `tab[,c(1,3)]`, `tab$transaction`. Que remarquez-vous? Quel est le nombre d'observations disponibles pour cette étude?
- Exécuter l'instruction `plot(tab)`. Que constate-t-on? Calculer la matrice de corrélations et commenter les résultats obtenus.

### Modèle multiple, MCO

On souhaite étudier le modèle de régression permettant d'expliquer le montant de la transaction immobilière en fonction des autres variables disponibles. Expliciter formellement le modèle attendu. La fonction `lsfit()` de R permet d'estimer les paramètres par MCO. Consulter l'aide cette fonction. Puis, Paramétrer de façon adéquate la fonction pour les données étudiées: `resmco=lsfit(x=..., y=...)`

- Exécuter l'instruction `summary(resmco)`. Détailler la structure de l'objet. Que contiennent les différents champs? Aider-vous de l'aide si besoin.
- Afficher dans la console les coefficients, $\hat{\beta}^T$ estimés du modèle . Quelle est la variation moyenne du montant d'achat du bien immobilier si on augmente la surface d'un $m^2$?
- Calculer la moyenne des erreurs quadratiques $E_2(\hat{\beta}) = \frac{1}{n}\sum_i \hat{\epsilon}_i^2$, puis $\sqrt{E_2(\hat{\beta})}$ en utilisant les champs proposés par la fonction `lsfit()` avec $\hat{\epsilon}_i = y_i - \hat{y}_i$ où $i$ indique une transaction $1 \leq i \leq n$. Conclusion.
- Afficher les erreurs relatives, $\epsilon_i^R$ entre la prédiction par le modèle et la valeur observée pour toutes les transactions $i$, avec $\epsilon_i^R = \frac{\hat{\epsilon}_i}{y_i} = \frac{y_i - \hat{y}_i}{y_i}$. Conclusion.
- Voyez-vous un lien (graphique) entre les erreurs du modèle et la valeur de la transaction? Conclusion.
- Calculer la valeur estimée par le modèle, , $\hat{y}_i = x_i\hat{\beta}^T$, pour chacune des transactions immobilières où $x_i$ représente les valeurs des variables explictives pour l'individu $i$. On pourra s'aider de l'instruction `as.matrix()` qui transforme un vecteur (une liste) en matrice.
- Visualiser sur un graphique le nuage de points des valeurs de la variable cible et de son estimation par le modèle. Indiquer sur le graphe la première bissectrice à l'aide de la fonction `abline()`. Que constatez-vous? .

### Modèle simple, MCO

Proposer un modèle de régression simple (une variable explicative) en justifiant votre choix? Donner les paramètres estimés.

- Afficher le montant des transactions en fonction de la variable retenue, ainsi que la droite de régression en vous aidant des fonctions R (`abline()`, `coef()`)
- Comparer les résultats actuels à ceux précédemment obtenus? Quel gain pratique est apporté par ce nouveau modèle?
- Un vendeur de biens immobiliers possèdent 5 appartements à vendre sur Paris de surface respective 155,178,200,220,250 $m^2$. En utilisant les données historiques déjà en votre possession, donner une estimation du montant de la transaction attendue pour chacun des biens.

### Validation croisée

*Cet exercice est à réaliser en fin de TP, après l'exercice suivant.* On souhaite à présent estimer les coefficients du modèle sur 75% des individus (base d'apprentissage), puis évaluer la qualité des résultats obtenus par ce premier modèle sur les 25% des individus restants (base de test). Utiliser la fonction `sample()` de R.

(1) Implémenter à l'aide d'une fonction R la liste d'instructions correspondantes. Calculer l'erreur quadratique moyenne de prédiction sur la base de test. (2) Calculer l'erreur quadratique moyenne en répétant cette procédure 10 fois. (3) Comparer l'erreur de prédiction à l'erreur résiduelle du modèle. Commenter les résultats obtenus.

## Régression et Pseudo-Inverse

Cette exercice a pour but d'illustrer l'impact de la dépendance entre variables explicatives sur l'estimation des coefficients de régression. Charger les données du fichier (Expseudo.txt) avec $Y$ variable cible (dernière colonne), $X_j$ variables explicatives, $1 \leq j \leq p$, observées pour $n$ observations. Spécifier les valeurs $n$ et $p$ ici définies, ainsi que le modèle de régression linéaire avec constante adapté à ce problème.
Partie A:

1. Calculer la matrice de variance-covariance empirique des variables explicatives. Que remarquez-vous?
2. Calculer la matrice $S = X^t X$ (on intégrera dans $X$ la place de la constante). Montrer que cette matrice n'est pas inversible et déterminer le rang de $S$ en utilisant une décomposition à valeurs singulières (`svd()`).
3. Donner la définition puis calculer matriciellement la pseudo-inverse de $S$, notée $S^{-1*}$.
4. Estimer à l'aide de la pseudo inverse les coefficients de régression avec constante, notées $\hat{\beta}_j^*$, $0 \leq j \leq p$.
5. Calculer les prédictions $\hat{Y}^P$ associées ainsi que l'erreur quadratique moyenne, notée $E^P$, entre les prédictions et les valeurs observées.
6. Déduire l'ensemble des solutions possibles de ce problème de MCO, notées $\hat{\beta}^\sharp$
7. Proposer une solution (simple) $\hat{\beta}^\sharp$ telle que $|\hat{\beta}^\sharp|_2^2 \simeq 10000$
8. Montrer que la solution proposée en partie A est de norme minimale.

Partie B:

1. Appliquer la commande `lsfit()` de R pour effectuer une régression linéaire sur les données précédentes. Donner la valeurs des coefficients estimés $\hat{\beta}_j^F$, $0 \leq j \leq p$.
2. Calculer les prédictions $\hat{Y}^F$ associées aux coefficients $\hat{\beta}^F$ et l'erreur quadratique moyenne, notée $E^F$, entre les prédictions et les valeurs observées
3. Comparer les résultats obtenues dans les parties A et B, en particulier les coefficients($\hat{\beta}^P$, $\hat{\beta}^F$), les prédictions ($\hat{Y}^P$, $\hat{Y}^F$), et les erreurs ($E^P$, $E^F$)

**Instructions R:** read.table(); cov(); cor(); det(); svd(), solve(); lm(); predict(), as.matrix(); as.matrix(); %*% pour le produit matriciel.