Premières notions de statistique
Régression Linéaire

Franck Picard

UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive

franck.picard@univ-lyon1.fr

Outline

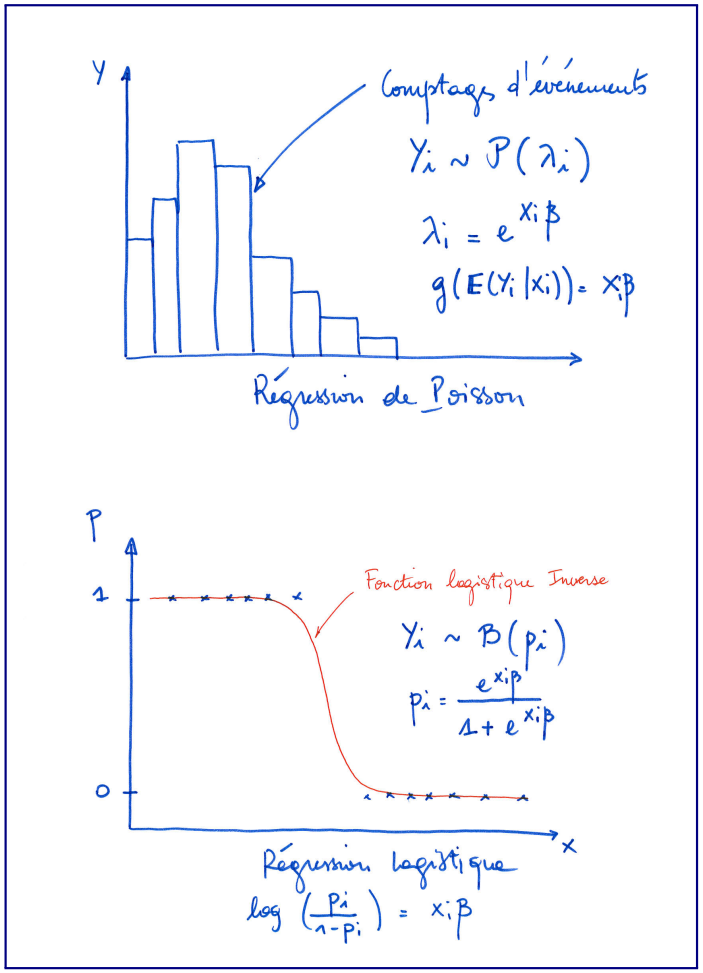
1. Principe généraux et typologie des modèles linéaires
2. Qu'est ce qu'un modèle de régression ?
3. Qu'est ce qu'un modèle de régression "linéaire" ?
4. Le modèle de régression linéaire simple
5. Tests, intervalles de confiance, et prédiction
6. Décomposition de la variance
7. Analyse des Résidus
8. Régression Linéaire Multiple

Préambule

- Une des stratégies les plus utilisée pour **planifier** des expériences et/ou **analyser** leurs résultats
- Les modèles linéaires permettent une modélisation "simple" des relations entre une variable **à expliquer**, souvent notée Y, et des **variables explicatives** souvent notées X (et souvent appelées covariables).
- Exemple: la taille des filles et des garcons est-elle la même ? le salaire dépend-il de l'age ? le médicament a-t-il un effet ? Le gène A prédispose-t-il à la maladie M ?
- Historiquement, le modèle linéaire a été développé par **Fisher**, avec applications en génétique et en agronomie

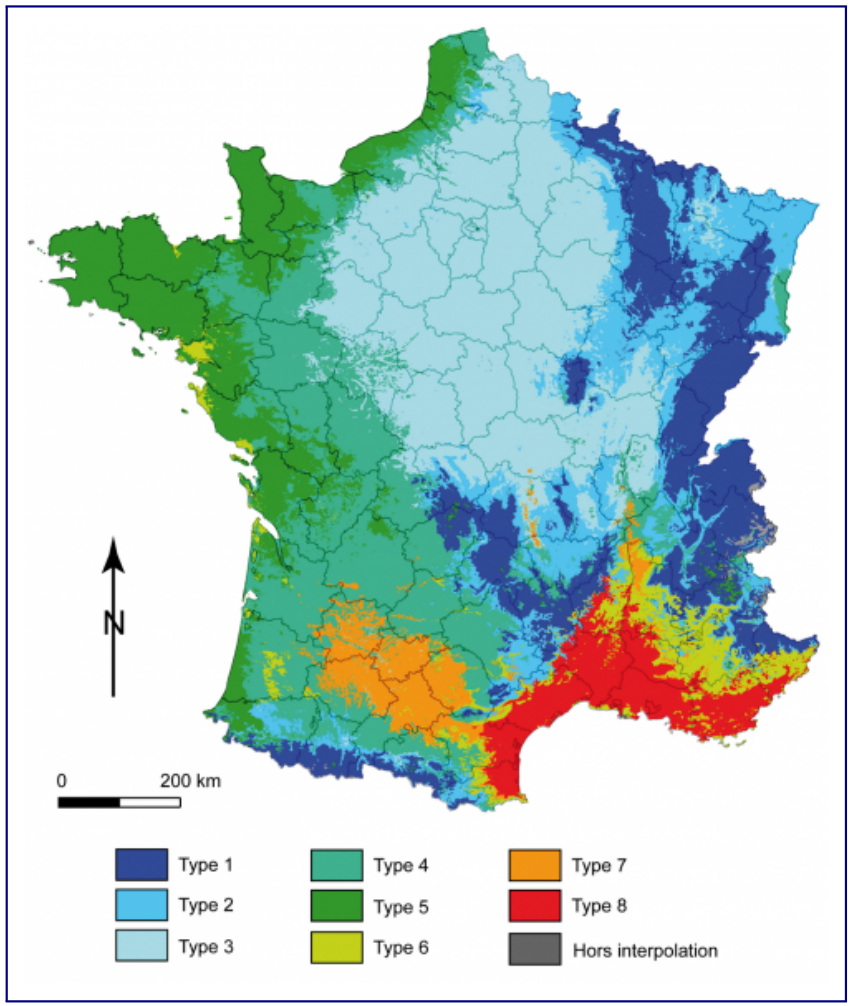
Modèle Linéaire Gaussien et Modèles Linéaires Généralisés

- Quelle **distribution** pour modéliser les observations ?
- Importance de l'analyse descriptive
- **Modèle linéaire gaussien** pour des observations pouvant être modélisées par une loi normale
- **Modèle linéaire généralisé** pour d'autres distributions (Poisson, Bernoulli...)



Modèles pour les observations (in)dépendantes

- Le modèle linéaire gaussien pour des observations qui **indépendantes**
- Séries chronologiques et modèles de **dépendance temporelle**
- statistique spatiale pour modéliser **dépendance spatiale**
- Les **modèles linéaires mixtes** permettent également de modéliser certains types de dépendance



Eléments de vocabulaire courant (et pertinent ?)

- L'ANOVA se caractérise par des variables explicatives discrètes ou catégorielles ou **qualitatives** (ex: Fille/Garcon, médicament A-B ou C)
- La Régression se caractérise par des variables explicatives continues ou **quantitatives** (ex: l'age, le poids)
- L'ANCOVA se caractérise par un mélange de variables qualitatives et quantitatives
- Il existe également des facteurs dits **ordinaux**: facteurs discrets ordonnés.

Ces trois modèles sont des modèles linéaires et se traitent de manière similaire: d'un point de vue **mathématique** et **pratique** il n'y a pas de différence

Outline

1 Principe généraux et typologie des modèles linéaires

2 Qu'est ce qu'un modèle de régression ?

3 Qu'est ce qu'un modèle de régression "linéaire" ?

4 Le modèle de régression linéaire simple

5 Tests, intervalles de confiance, et prédiction

6 Décomposition de la variance

7 Analyse des Résidus

8 Régression Linéaire Multiple

Premières notations

- On suppose que l'on dispose de n observations $(y_1, \ldots, y_n)$ que l'on modélise par des variables aléatoires gaussiennes **indépendantes** $(Y_1, \ldots, Y_n)$: $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
- On suppose que la variance de toutes les observations est la même : c'est l'hypothèse d'**homoscédasticité** (σ^2 **est constante**).
- On observe également des covariables $(x_1, \ldots, x_n)$, **sur les mêmes individus**. Les données dont on dispose sont en fait les **couples** $(y_i, x_i)_i$.
- Exemples : le poids y_i d'une personne i et sa taille x_i, le rendement d'une culture y_i et la dose d'engrais x_i.

Pour un modèle linéaire "standard" on suppose que les Y_i sont **aléatoires** et que les x_i sont **fixées**

Notion d'espérance conditionnelle

- Une stratégie de modélisation pour étudier les relations entre y_i et x_i est de supposer que les covariables ont une influence sur **l'espérance des** Y_i
- On modélise l'espérance de Y_i **conditionnellement** aux valeurs observées des X_i à x_i:

$$\begin{aligned} Y_i|\{X_i = x_i\} &\sim \mathcal{N}(\mu(x_i), \sigma^2) \\ \mu(x_i) &= \mathbb{E}(Y_i|X_i = x_i) = \int y_i f_{Y|X}(y_i; x_i) dy \end{aligned}$$

- $\mu(x_i)$ s'appelle la **fonction de régression**: c'est la fonction qui relie les x_i aux observations.
- Ce que l'on néglige en considérant l'espérance conditionnelle, c'est la variabilité des covariables que l'on suppose fixées.

Et la variance conditionnelle ?

- Qu'en est-il de la relation entre les covariables et la variance des Y ?
- On note $\mathbb{V}(Y_i|X_i = x_i)$ cette variance conditionnelle

$$\mathbb{V}(Y_i|X_i = x_i) = \mathbb{E}(Y_i^2|X_i = x_i) - \mathbb{E}^2(Y_i|X_i = x_i)$$

Dans le modèle linéaire gaussien on suppose que la variabilité des observations Y_i ne dépend pas des covariables

$$\mathbb{V}(Y_i|X_i = x_i) = \sigma^2$$

- Exemple: la variabilité de la taille des filles est la même que la variabilité de la taille des garcons.
- Ce n'est pas forcément une hypothèse réaliste, mais elle permet de faire les calculs
- Il existe des stratégies pour "stabiliser" la variance (méthode delta)

Définition des variables résiduelles

- Jusqu'à présent, le modèle s'écrivait : $Y_i|\{X_i = x_i\} \sim \mathcal{N}(\mu(x_i), \sigma^2)$
- On peut considérer la nouvelle variable

$$\varepsilon_i = Y_i - \mathbb{E}(Y_i|X_i = x_i) \sim \mathcal{N}(0, \sigma^2)$$

- C'est l'écart entre l'observation Y_i et son espérance conditionnelle.
- ε_i est **résidu aléatoire**: c'est **erreur aléatoire** que l'on commettrait en remplacant Y_i par $\mu(x_i)$.
- On propose une autre écriture du modèle linéaire gaussien:

$$Y_i = \mu(x_i) + \varepsilon_i, \ \varepsilon_i \underset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Le paramètre σ^2 s'interprète comme la **variabilité des erreurs aléatoires**

Outline

1 Principe généraux et typologie des modèles linéaires

2 Qu'est ce qu'un modèle de régression ?

3 Qu'est ce qu'un modèle de régression "linéaire" ?

4 Le modèle de régression linéaire simple

5 Tests, intervalles de confiance, et prédiction

6 Décomposition de la variance

7 Analyse des Résidus

8 Régression Linéaire Multiple

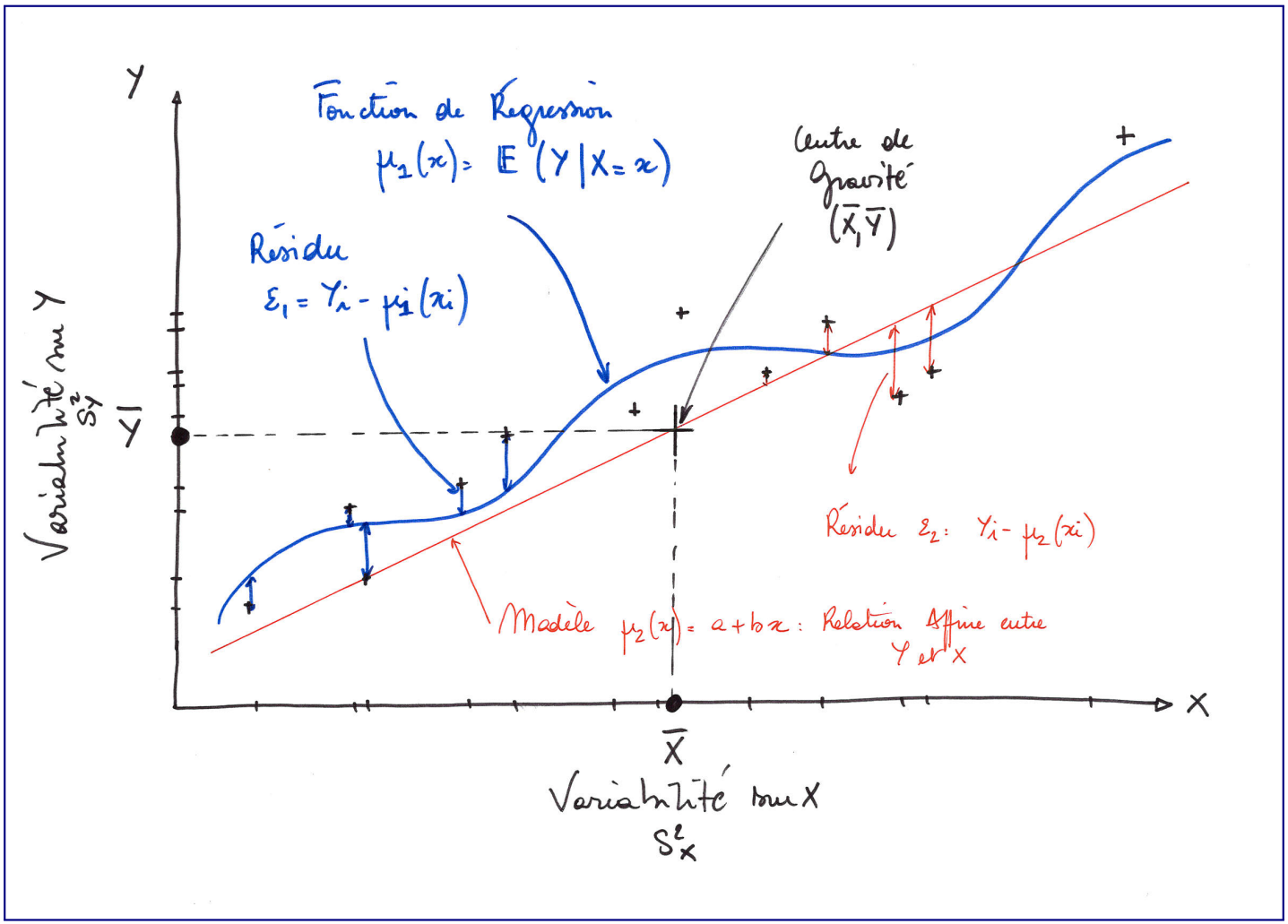
Quelle forme pour la fonction de régression ?

- Le modèle $\mathbb{E}(Y_i|X_i = x_i) = \mu(x_i)$ est très général, et on ne connait pas forcément la forme de la fonction μ
- Le cadre de la **régression fonctionnelle** s'intéresse à l'estimation de la fonction μ directement
- Dans le modèle linéaire, on fait des **hypothèses supplémentaires** sur la forme de μ:
  1. On suppose que μ dépend de **paramètres** $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$. p représente le nombre de covariables disponibles
  2. On suppose que μ_β est une **fonction affine**

Dans un modèle linéaire on supposera que

$\mu(x_i) = \beta_0 + \beta_1 x_i$, et que (β_0, β_1) sont **fixes mais inconnus**

Illustration de la Régression Linéaire



Variables et Paramètres

- En faisant l'hypothèse que $\mu(x_i) = \beta_0 + \beta_1 x_i$, on a reformulé le problème de l'explication de Y_i par x_i.
- En choisissant la forme de la fonction de régression, on peut se focaliser sur les paramètres (β_0, β_1)
- Etudier les liens entre Y et x revient désormais à étudier les paramètres du modèle (estimation, test).
- Si $\beta_1 = 0$, alors on supposera que x n'a pas d'influence sur Y

On interprétera β_1 comme **l'effet de la covariable x sur la réponse Y**.

Regression Linéaire / Regression Non Linéaire

- La linéarité du modèle linéaire concerne les **paramètres** et pas forcément les **variables**
- Exemples :$\mu(x) = \beta_0 + \beta_1 \cos(x)$ est une fonction linéaire en β_0, β_1 mais pas en x, $\mu(x) = x \exp(\beta_0)/(\beta_0 + \beta_1)$ n'est pas linéaire en les paramètres, mais elle est linéaire en x
- Dans certaines situations on peut se ramener à un modèle linéaire par transformations. Exemples
  - $\mu(x) = \beta_0 \exp(\beta_1 x)$
  - $\mu(x) = \beta_0 x^{\beta_1}$
  - $\mu(x) = \beta_0 + \beta_1/x$

Attention aux transformations ! Il faut adapter l'interprétation des paramètres !

Régression Linéaire Simple / Régression Linéaire Multiple

- La régression **linéaire simple** consiste à étudier la relation affine entre Y et un seul régresseur x
- La régression **linéaire multiple** s'intéresse aux relations entre Y_i et plusieurs régresseurs
- On notera x_j le régresseur j, et x_{ij} son observation sur l'individu i.
- On associe à x_j le paramètre β_j commun à tous les $(x_{1j}, \ldots, x_{nj})$.
- La fonction de régression μ dépend de p régresseurs $(x_1, \ldots, x_p)$:

$$\mu(x_1, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

β_j s'interprétera comme l'effet de la covariable x_j.

Régresion / ANOVA, même combat !

- On considère souvent que la régression se limite au cas où x est quantitatif, mais elle peut se généraliser au cas ou x est discret
- Si est une variable discrète : $x_i = 1$ si l'individu i est un garcon, 0 sinon.
- On utilise la notation $1_A = 1$ si l'événement A est réalisé, 0 sinon
- On s'intéresse au poids Y_i des individus en fonction de leur genre, et on peut définir la fonction de régression suivante:

$$\mu(x) = \beta_0 + \beta_1 1_{\{x=1\}} + \beta_2 1_{\{x=0\}}$$

La régression peut donc considérer des facteurs quantitatifs **ET** qualitatifs. Par contre, elle est contrainte ce que la distribution de la réponse soit gaussienne.

Outline

1 Principe généraux et typologie des modèles linéaires

2 Qu'est ce qu'un modèle de régression ?

3 Qu'est ce qu'un modèle de régression "linéaire" ?

4 Le modèle de régression linéaire simple

5 Tests, intervalles de confiance, et prédiction

6 Décomposition de la variance

7 Analyse des Résidus

8 Régression Linéaire Multiple

Contexte et objectifs - 1

- On observe 2 caractéristiques quantitatives y et x sur une population de n individus. Les données sont donc sous la forme de couples $(y_i, x_i)_i$.
- On suppose qu'il existe une **relation affine** entre y et x, qui dépend de deux paramètres (β_0, β_1):

$$\mu(x) = \beta_0 + \beta_1 x$$

- On suppose également que les observations sont des réalisations de variables aléatoires gaussiennes i.i.d, telles que

$$Y_i | X_i = x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

- On introduit les variables d'erreur **aléatoires** ε_i, indépendantes, gaussiennes centrées de variance σ^2 **constante**

Contexte et objectifs - 2

- Le modèle de régression simple s'écrit:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$

- β_0 représente la valeur moyenne des observations Y_i quand $x_i = 0$ (interprétation ?)
- β_1 représente la **pente de la droite de régression**, et correspond à la variation moyenne de Y si x augmentait d'une unité, **et si la vraie relation entre Y et x était linéaire**.
- Objectifs de l'étude statistique:
  - Estimer les paramètres du modèle (β_0, β_1) et σ^2
  - Etudier la pertinence du modèle: analyse des résidus, tests
  - Construire un intervalle de confiance et de prédiction de la droite de régression

Estimation des paramètres par la méthode des moindre-carrés

- A partir des observations (Y_i, x_i) on souhaite trouver des estimateurs $\widehat{\beta}_0, \widehat{\beta}_1$.
- La stratégie communément envisagée est celle des moindre-carrés ordinaires (MCO)
- On considère l'erreur quadratique moyenne EQM définie par:

$$\begin{aligned} \text{EQM}(\mu) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \\ \text{EQM}(\beta_0, \beta_1) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - [\beta_0 + \beta_1 x_i])^2 \end{aligned}$$

L'EQM quantifie la distance entre le modèle $\mu(x_i)$ et les observations Y_i ou encore la variabilité des erreurs ε_i

Système d'équations

- $(\widehat{\beta}_0^{\text{MCO}}, \widehat{\beta}_1^{\text{MCO}})$ sont les solutions de

$$\frac{\partial \text{EQM}(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \frac{\partial \text{EQM}(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

- Les dérivées partielles sont:

$$\begin{aligned} \frac{\partial \text{EQM}(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - [\beta_0 + \beta_1 x_i]) \\ \frac{\partial \text{EQM}(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (Y_i - [\beta_0 + \beta_1 x_i]) \end{aligned}$$

Estimateurs des moindre-carrés ordinaires

- On utilise les notations usuelles:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Les estimateurs des MCO sont alors:

$$\begin{aligned} \widehat{\beta}_0^{\text{MCO}} &= \overline{Y} - \widehat{\beta}_1^{\text{MCO}}\overline{x} \\ \widehat{\beta}_1^{\text{MCO}} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{aligned}$$

- On appelle **droite de régression** la droite d'équation

$$y = \widehat{\beta}_0^{\text{MCO}} + \widehat{\beta}_1^{\text{MCO}} x \textbf{ pour tout } (y, x)$$

Interprétation des estimateurs des paramètres

- $\widehat{\beta_0}$ est l'estimateur de l'ordonnée à l'origine de la droite de régression
- Ce paramètre n'est pas toujours interprétable ! (dépend de la signification de x et du fait que x soit centrée ou non)
- Autre écriture:

$$\overline{Y} = \widehat{\beta_0} + \widehat{\beta_1}\overline{x}$$

- La droite de régression passe par le **centre de gravité** du nuage de points $(\overline{Y}, \overline{x})$.
- Précaution : la technique des MCO créée des estimateurs sensibles aux valeurs atypiques (cf. analyse des résidus)

Covariance théorique entre deux variables aléatoires

- La covariance et la corrélation théoriques entre deux variables aléatoires (X, Y) sont définies par:

$$\text{Cov}(Y, X) = \mathbb{E}([Y - \mathbb{E}Y][X - \mathbb{E}X]) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

$$\text{Corr}(Y, X) = \frac{\text{Cov}(Y, X)}{\sqrt{\mathbb{V}(Y)\mathbb{V}(X)}}$$

- Si Y était de la forme $\beta_0 + \beta_1 X$, alors:

$$\text{Cov}(Y, X) = \text{Cov}(\beta_0 + \beta_1 X, X) = \beta_1 \mathbb{V}(X)$$

$$\text{Corr}(Y, X) = \text{Corr}(\beta_0 + \beta_1 X, X) = 1$$

La covariance entre deux variable mesure **la part de dépendance linéaire** entre X et Y. La corrélation est un coefficient **sans unité**, c'est la version standardisée de la covariance

Interprétation géométrique de la corrélation

- Si $\text{Corr}(X, Y) = 0$ alors il n'y a pas de relation linéaire entre Y et X.
- Si $|\text{Corr}(X, Y)| = 1$ alors la connaissance de X détermine exactement celle de Y
- Si $\text{Corr}(X, Y) > 0$ alors quand X augmente, Y augmente en moyenne
- Si $\text{Corr}(X, Y) < 0$ alors quand X augmente, Y diminue en moyenne (anticorrélation)

La corrélation n'informe en rien sur la causalité entre X et Y mais permet de détecter un lien de type linéaire

Covariance empirique entre deux variables aléatoires

- La covariance et la corrélation empiriques sont définies par:

$$\widehat{\text{Cov}}(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2, \quad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\widehat{\text{Corr}}(Y,X) = \frac{\widehat{\text{Cov}}(Y,X)}{\sqrt{S_Y^2 S_X^2}}$$

- Ce sont des **estimateurs sans biais** de la covariance et du coefficient de corrélation

Interprétation de l'estimateur du coefficient de régression

- L'estimateur de β_1 peut s'écrire:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\widehat{\mathrm{Cov}}(Y, X)}{S_X^2} = \widehat{\mathrm{Corr}}(Y, X)\frac{S_Y}{S_X}$$

- On peut s'intéresser à l'EQM en le point $(\widehat{\beta}_0, \widehat{\beta}_1)$:

$$\mathrm{EQM}(\widehat{\beta}_0, \widehat{\beta}_1) = S_Y^2(1 - \widehat{\mathrm{Corr}}(Y, X)) = \frac{n-2}{n}\widehat{\sigma}^2$$

- L'erreur quadratique minimale est d'autant plus faible que la corrélation entre X et Y est forte.
- L'erreur quadratique minimale sert d'estimateur à la variance des résidus (correction pour le biais)

Outline

1 Principe généraux et typologie des modèles linéaires

2 Qu'est ce qu'un modèle de régression ?

3 Qu'est ce qu'un modèle de régression "linéaire" ?

4 Le modèle de régression linéaire simple

5 Tests, intervalles de confiance, et prédiction

6 Décomposition de la variance

7 Analyse des Résidus

8 Régression Linéaire Multiple

Avec ou sans modèle de distribution ?

- Il est important de noter que la construction du modèle de régression et l'estimation des paramètres par MCO ne fait pas appel aux hypothèses de distribution
- Les hypothèses de distribution sont essentielles lorsqu'il s'agit de construire des tests et des intervalles de confiance et de prédiction
- Hypothèses fondamentales:
  - Les observations sont **indépendantes**
  - La variance des erreurs est **constante** σ^2
  - La loi des erreurs est une **loi normale** $\mathcal{N}(0, \sigma^2)$

Loi des estimateurs du modèle linéaire gaussien

- On admettra que les estimateurs des paramètres β_j sont gaussiens sans biais, tels que

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\mathbb{V}(\widehat{\beta}_j)}} \sim \mathcal{N}(0,1)$$

- $\mathbb{V}(\widehat{\beta}_j)$ est la variance de l'estimateur du paramètre β_j et on admettra également qu'elle est de la forme:

$$\mathbb{V}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\overline{x})^2}{(n-1)S_{X_1}^2} \right), \quad \mathbb{V}(\widehat{\beta}_1) = \frac{\sigma^2}{(n-1)S_{X_1}^2}$$

- $S_{X_1}^2$ représente la dispersion de X_1
- σ^2 peut être estimée par l'EQM en $(\widehat{\beta}_0, \widehat{\beta}_1)$, d'où

$$\widehat{\mathbb{V}(\widehat{\beta}_1)} = \frac{\widehat{\sigma}^2}{(n-1)S_{X_1}^2} \sim \chi^2(n-2)$$

Test des paramètres et intervalles de confiance

- On se pose la question de l'effet de la covariable $H_0 : \{\beta_1 = 0\}$
- D'après les hypothèses et les propriétés précédentes, on peut construire un test à partir de la statistique:

$$\frac{\widehat{\beta}_j - 0}{\sqrt{\widehat{\mathbb{V}(\widehat{\beta}_j)}}} \underset{H_0}{\sim} \mathcal{T}(n-2)$$

- De même, on peut construire un intervalle de confiance du paramètre de la pente:

$$\text{IC}_{1-\alpha}(\beta_1) = \left[\widehat{\beta}_1 \pm t^{n-2}_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2}{(n-1)S^2_{X_1}}}\right]$$

Intervalle de confiance de la droite de régression

- On a construit des intervalles de confiance pour les paramètres β_0 et β_1.
- On peut également construire un intervalle de confiance pour la droite de régression:

$$\begin{aligned}
IC_{1-\alpha}(\beta_0) &= \left[\widehat{\beta}_0 \pm t^{n-2}_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n}+\frac{(\overline{x})^2}{(n-1)S^2_{X_1}}\right)}\right] \\
IC_{1-\alpha}(\beta_1) &= \left[\widehat{\beta}_1 \pm t^{n-2}_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2}{(n-1)S^2_{X_1}}}\right] \\
IC_{1-\alpha}(\beta_0+\beta_1 x_i) &= \left[\widehat{\beta}_0+\widehat{\beta}_1 x_i \pm t^{n-2}_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n}+\frac{(x_i-\overline{x})^2}{(n-1)S^2_{X_1}}\right)}\right]
\end{aligned}$$

- La largeur de l'intervalle de confiance dépend de la distance $(x_i - \overline{x})^2$

Interprétation du test de la pente

- Il faut toujours se rappeler que la régression linéaire ne s'intéresse qu'à la part linéaire de la dépendance entre deux variables
- Si on rejette H_0, cela ne signifie pas que tous les liens entre les deux variables sont captés par le modèle.
- Si H_0 n'est pas rejetée:
  - il n'existe pas de lien (du tout) entre les variables
  - il n'y a pas suffisamment de données pour détecter ce lien (pb de puissance)
  - le lien entre les variables n'est pas de type linéaire

Prédiction

- On peut considérer l'exercice de régression en deux temps: apprentissage et prédiction
- **Apprentissage**: On considère $(Y_i, x_i)_i$ un échantillon d'apprentissage sur lequel on apprend la forme de $\mu(x)$ et on en propose une :

$$\widehat{\mu}(x) = \widehat{\beta}_0\left((Y_i, x_i)_{i=1,n}\right) + \widehat{\beta}_1\left((Y_i, x_i)_{i=1,n}\right) x$$

- **Prédiction**: pour un nouvel x_0 est ce que l'on peut prédire la réponse Y_0 ?

$$\widehat{Y}_0 = \widehat{\mu}(x_0) = \widehat{\beta}_0\left((Y_i, x_i)_{i=1,n}\right) + \widehat{\beta}_1\left((Y_i, x_i)_{i=1,n}\right) x_0$$

Variance de prédiction

- Quelle est l'erreur que l'on commettrait en prédisant un nouvel Y_0 par $\widehat{\mu}(x_0)$
- Un point essentiel est que pour cet x_0 le résidu ε_0 de variance σ^2 n'est pas observé:

$$\widehat{Y}_0 = \widehat{\beta}_0\left((Y_i, x_i)_{i=1,n}\right) + \widehat{\beta}_1\left((Y_i, x_i)_{i=1,n}\right) x_0 + \varepsilon_0$$

- Ce résidu n'ayant pas été observé, il est indépendant de l'échantillon d'apprentissage:

$$\begin{aligned}
\mathbb{V}(\widehat{Y}_0) &= \mathbb{V}\left[\widehat{\beta}_0\left((Y_i, x_i)_{i=1,n}\right) + \widehat{\beta}_1\left((Y_i, x_i)_{i=1,n}\right) x_0\right] + \mathbb{V}(\varepsilon_0) \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)S_{X_1}^2}\right) + \sigma^2
\end{aligned}$$

Intervalle de prédiction

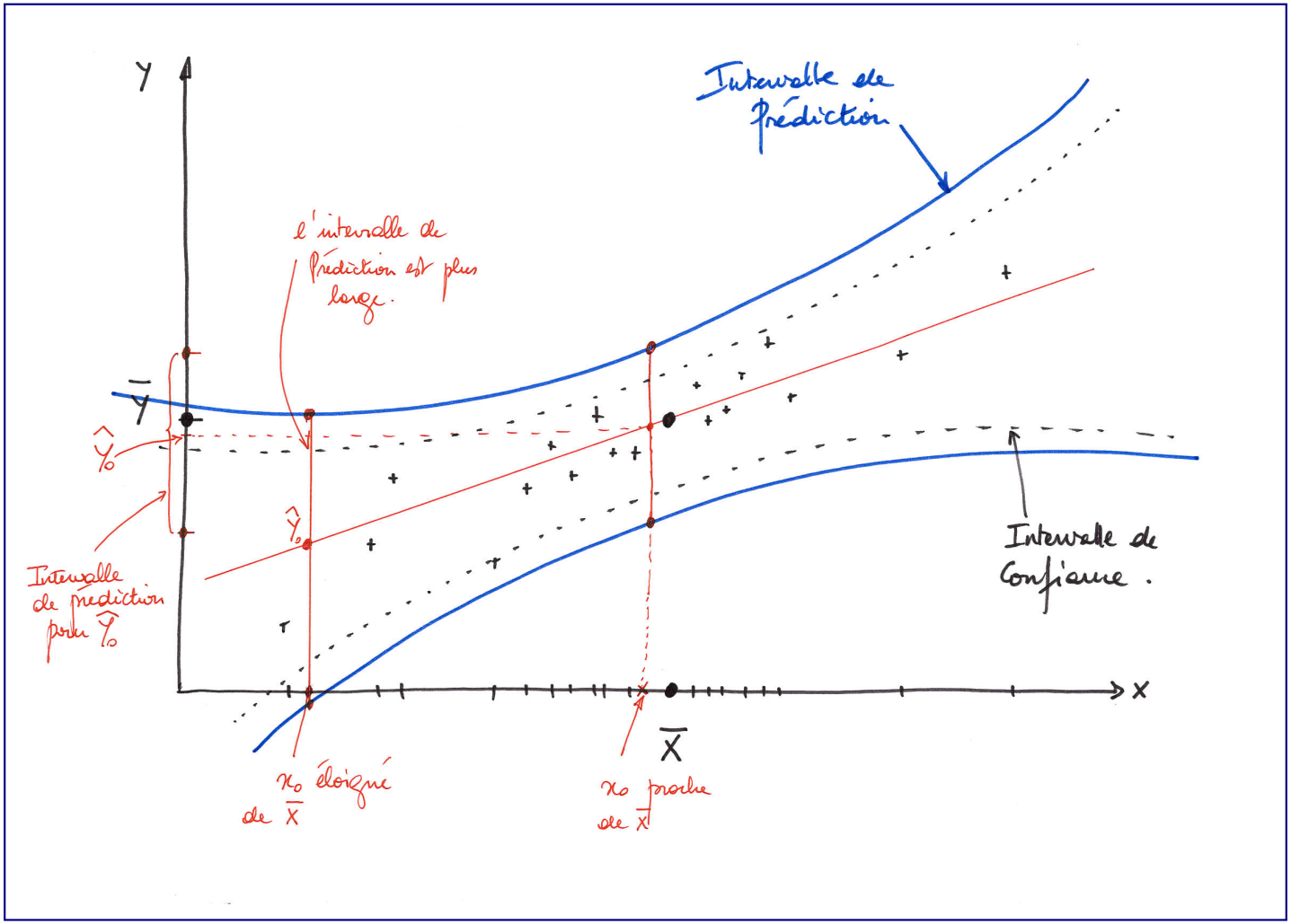
- La non observation de x_0 se traduit par un terme σ^2 supplémentaire
- L'intervalle de prédiction est plus large que l'intervalle de confiance

$$IC_{1-\alpha}(\beta_0 + \beta_1 x_i) = \left[\widehat{\beta}_0 + \widehat{\beta}_1 x_i \pm t^{n-2}_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{(n-1)S^2_{X_1}}\right)}\right]$$

$$IP_{1-\alpha}(\beta_0+\beta_1 x_0) = \left[\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm t^{n-2}_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)S^2_{X_1}}\right)}\right]$$

- La qualité de la prédiction dépend elle aussi de la distance au centre de gravité $(x_0 - \overline{x})^2$

Illustration des Intervalles de Confiance/Prédiction



Outline

1 Principe généraux et typologie des modèles linéaires

2 Qu'est ce qu'un modèle de régression ?

3 Qu'est ce qu'un modèle de régression "linéaire" ?

4 Le modèle de régression linéaire simple

5 Tests, intervalles de confiance, et prédiction

6 Décomposition de la variance

7 Analyse des Résidus

8 Régression Linéaire Multiple

Le cas particulier de la régression linéaire simple

- Le modèle est le suivant: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (+hypothèses)
- Les sommes de carrés : $\widehat{Y}_i(x_i) = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$:

$$\begin{aligned} \text{SCT}(\mathbf{Y}) &= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \\ \text{SCM}(\mathbf{Y}, \mathbf{X}) &= \sum_{i=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \overline{Y})^2 \\ \text{SCR}(\mathbf{Y}, \mathbf{X}) &= \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2 \end{aligned}$$

Le cas particulier de la régression linéaire simple

- La table d'Analyse de la variance s'écrit:

Source	df	SS	MS	F
Model	1	SCM	SCM/1	$\frac{SCM/1}{SCR/(n-2)}$
Error	n-2	SCR	SCR/(n-2)	
Total	n-1	SCT	SCT/(n-1)	

- Le test de Fisher consiste à comparer le modèle nul au modèle complet
- Dans le cas de la régression simple, il revient à tester l'hypothèse $\mathcal{H}_0\{\beta_1 = 0\}$

Coefficient de corrélation et de détermination

- Une confusion est souvent faite entre l'interprétation de R^2 dans le cas de la régression simple et dans le cas général.
- Dans le cas de la régression simple, il s'écrit: $R^2 = \widehat{\beta}_1^2 S_X^2 / S_Y^2$
- $R^2 \simeq 1$ indique que le coefficient de corrélation empirique entre les observations et la covariable est proche de 1, donc que la modélisation des observations par une droite est très satisfaisante

Dans le cas général, R^2 n'est pas un coefficient de corrélation

Outline

1 Principe généraux et typologie des modèles linéaires

2 Qu'est ce qu'un modèle de régression ?

3 Qu'est ce qu'un modèle de régression "linéaire" ?

4 Le modèle de régression linéaire simple

5 Tests, intervalles de confiance, et prédiction

6 Décomposition de la variance

7 Analyse des Résidus

8 Régression Linéaire Multiple

Motivation et définition

- La première question à se poser avant de regarder les résultats du modèle : les hypothèses qui ont été faites au départ sont-elles respectées ?
- Hypothèses fondamentales:
  - (ε_i) sont gaussien $\mathcal{N}(0, \sigma^2)$
  - (ε_i) sont indépendants
  - σ^2 est constante (ne varie pas avec x)
- Malheureusement, les valeurs exactes des résidus resteront inconnues, mais on les estimera par :

$$\widehat{\varepsilon_i} \quad = \quad Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)$$

- On introduit le coefficient h_{ii} tel que:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Variance des résidus estimés

Même sous l'hypothèse d'homoscédasticité les résidus estimés n'ont pas la même variance ! (mais ils sont centrés par construction)

$$\begin{aligned} h_{ii} &= \frac{1}{n}\left(1 + \frac{(x_i - \overline{x})^2}{S_X^2}\right) \\ S_X^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ \mathbb{V}(\widehat{\varepsilon}_i) &= \sigma^2(1 - h_{ii}) \end{aligned}$$

Leur dispersion au point (i) dépend de la distance de x_i au centre de gravité $\overline{x}$.

Résidus estimés "réduits"

- On peut réduire les résidus estimés en considérant la variable

$$r_i = \frac{\widehat{\varepsilon}_i}{\mathbb{V}\left(\widehat{\varepsilon}_i\right)} \sim \mathcal{N}(0,1)$$

- Mais on ne connait pas $\mathbb{V}\left(\widehat{\varepsilon}_i\right)$ que l'on estime par:

$$\widehat{\mathbb{V}\left(\widehat{\varepsilon}_i\right)} = \widehat{\sigma}^2(1 - h_{ii})$$

- Les résidus estimés "réduits" suivent une loi de Student à (n-2) degrés de liberté:

$$\frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\mathbb{V}\left(\widehat{\varepsilon}_i\right)}}} \sim \mathcal{T}_{n-2}$$

Influence de la i^{eme} observation

- Le critère des moindre carrés est très sensible aux valeurs aberrantes (loin du centre de gravité du nuage $(\overline{Y}, \overline{x})$).
- L'étude des résidus se fait également en étudiant l'influence des poins aberrants et la stabilité des estimations
- On procède en enlevant l'observation (i) et on calcule $\widehat{\sigma}^2_{(i)}$, l'estimateur de la variance des résidus calculés sur $(n-1)$ observations en se privant de la ième.
- On définit les résidus "studentisés" :

$$t_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \sim \mathcal{T}_{n-3}$$

- L'analyse graphique consiste à explorer la distribution des t_i.

Effet Levier, Distance de Cook

- Le terme h_{ii} représente le poids de l'observation i sur sa propre estimation. On peut montrer que:

$$\begin{aligned} \widehat{Y}_i &= h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j \\ h_{ij} &= \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum_{h=1}^{n} (x_h - \overline{x})^2} \end{aligned}$$

- Si h_{ii} est grand ($\geq 1/2$), alors le point i est un point levier (point atypique)
- La distance de Cook est utilisée pour mesurer l'influence de l'observation i sur l'estimation:

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{Y}_{(i)j} - \widehat{Y}_j)^2}{2\widehat{\sigma}^2} = \frac{h_{ii}}{2(1 - h_{ii})} r_i^2$$

Le graph des résidus

- Par le théorème de Cochran on sait que $\widehat{Y}_i$ et $Y_i - \widehat{Y}_i$ sont indépendants
- Pour vérifier cette hypothèse, on trace le graphe des résidus $Y_i - \widehat{Y}_i$ vs $\widehat{Y}_i$
- Ce graph ne doit montrer aucune tendance, et doit être centré en zéro
- Il permet de vérifier visuellement l'hypothèse d'homoscédasticité.
- C'est le premier indicateur à regarder:

si le graph des résidus n'est pas correctement structuré, alors les résultats du modèle n'ont aucun sens car les hypothèses ne sont pas respectées.

Outline

1 Principe généraux et typologie des modèles linéaires

2 Qu'est ce qu'un modèle de régression ?

3 Qu'est ce qu'un modèle de régression "linéaire" ?

4 Le modèle de régression linéaire simple

5 Tests, intervalles de confiance, et prédiction

6 Décomposition de la variance

7 Analyse des Résidus

8 Régression Linéaire Multiple

Motivations, Définitions

- En pratique, on souhaite souvent expliquer les variations d'une réponse à l'aide de **plusieurs** covariables
- On note Y_i la réponse, et $\mathbf{x}_i = (x_{0i}, x_{i1}, \ldots, x_{ip})$ le vecteur des covariables de taille $p+1$ (convention $x_{0i} = 1$).
- On peut considérer des puissances successives d'une variable: c'est un modèle polynomial $\mathbb{E}[Y_i|x_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$
- On peut considérer des variables différentes $\mathbb{E}[Y_i|x_i, z_i, x_i] = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 w_i$
- La méthode d'estimation sera identique (estimateurs des moindre-carrés)
- Mais la question de la confusion d'effet devient importante : les covariables ont-elles des liens entre elles ? Sont-elles corrélées ? Peut-on distinguer l'effet d'une covariables sachant tout le reste ?

Régression linéaire multiple, notations matricielles

- Le modèle de régression linéaire multiple à p régresseurs s'écrit:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- On considère les vecteurs $\mathbf{Y} = [Y_1, \ldots, Y_n]^T$ et $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^T$ de taille n,
- On considère la matrice $\mathbf{X}$ de taille $n \times p$, telle que la colonne j de $\mathbf{X}$ correspond à la covariable $\mathbf{x}_j$ pour tout i et la ligne i correspond à l'enregistrement de toutes les variables pour l'observation i.
- On considère $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]^T$ le vecteur des coefficients $(p \times 1)$.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$$

Estimation

- L'estimation se déroule comme dans le cas de la régression simple (p=1), avec le critère des Moindre-carrés:

$$MC(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} (Y_i - [\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}])^2$$

- Un des critères important pour pouvoir résoudre le système est qu'il n'existe pas de redondance dans les covariables ($rang(\mathbf{X}) = p + 1$).
- Un estimateur de la variance résiduelle est donné par:

$$\widehat{\sigma}^2 = \frac{1}{n - p - 1} MC(\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$$

- Un prédicteur de Y_i sera donné par: $\widehat{Y}_i = \sum_{j=0}^{p} \widehat{\beta}_j x_{ij}$.
- La somme des carrés du modèle s'écrit :

$$SCM = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$$

Table d'analyse de la variance

Source	df	SS	MS	F
Model	p	SCM	SCM/p	$\frac{SCM/p}{SCR/(n-p-1)}$
Error	n-p-1	SCR	SCR/(n-p-1)	
Total	n-1	SCT	SCT/(n-1)	

- Le modèle peut être testé globalement $H_0 : \{\beta_0 = \beta_1 = \ldots = \beta_p\}$

Mais cette stratégie globale n'est pas très informative car elle revient à comparer le modèle nul au modèle complet et H_0 sera rejetée dès qu'un seul des β_j est non nul

Qualité d'ajustement et taille du modèle

- On peut définir la taille d'un modèle par le nombre de **paramètres libres** qui le caractérisent
- Dans le cas de la régression multiple: 1 pour la moyenne générale (β_0), p pour les $(\beta_j)_j$ et 1 pour la variance des erreurs σ^2

La somme des carrés totale étant constante quand la taille du modèle augmente, SCM augmente et SCE diminue

- Plus on ajoute des variables, plus le modèle s'ajuste aux données, mais plus on commet d'erreurs d'estimation.
- Si on s'intéresse uniquement au $R^2 = SCM/SCT = 1 - SCR/SCT$, il est croissant avec le nombre de paramètres.

La sélection de variables

- La pratique de la modélisation suppose un équilibre entre:
  - Un grand nombre de variables explicatives, pour avoir un modèle "exhaustif" qui prend en compte une certaine "complexité" des données
  - Un nombre raisonnable de paramètres (interprétabilité, parcimonie)
- La sélection de variables consiste à choisir un sous ensemble de q variables parmi les p disponibles (un sous modèle)

$$SCT = SCM_p + SCR_p = SCM_q + SCR_q$$

- Si les q variables sélectionnés sont pertinentes, alors on suppose que les $p - q$ restantes on un effet négligeable sur la somme des carrés résiduelle

R^2 ajusté

Le coefficient de détermination R^2 étant croissant avec le nombre de variables, il ne peut être utilisé que pour comparer des modèles ayant le même nombre de paramètres

- On définit le R^2 ajusté pour comparer des modèles de tailles différentes

$$R^2_{\text{aj}} = 1 - \frac{SCR_p/(n-p-1)}{SCT/(n-1)}$$

- le R^2_{aj} compare les sommes des carrés moyennes (ajustées au nombre de paramètres)
- Il existe d'autres critères pour comparer des modèles entre eux (C_p de Mallows, AIC, BIC).

Algorithmes de sélection

- Il existe 2^p modèles différents quand on considère p variables
- Quand p est grand, on ne peut pas tous les explorer.
- Il existe plusieurs stratégies pour explorer les modèles:
  - Sélection Forward: une variable est ajoutée à chaque pas
  - Sélection Backward: une variable est enlevée à chaque pas
  - Sélection stepwise: introduction de variables supplémentaires, mais élimination des redondantes à chaque étape
- Le critère de choix du modèle est souvent défini à part (AIC, BIC, Cp, R2aj) et utilisé dans l'algorithme