Master 1 Informatique
Éléments de statistique inférentielle

Faicel Chamroukhi
Maître de Conférences
UTLN, LSIS UMR CNRS 7296

email: chamroukhi@univ-tln.fr
web: chamroukhi.univ-tln.fr

2014/2015

Plan I

1. Méthode des Moindres Carrées

2. Régression linéaire

Méthode des Moindres Carrées (Least Squares)

- Définition des Moindres Carrés
- Moindres Carrés
- Propriétés de l'estimateur des moindres carrés

Méthode des Moindres Carrées I

La méthode des moindres carrés consiste à estimer les paramètres d'un modèle en minimisant les écarts quadratiques entre les données observées, d'une part, et leurs valeurs attendues, d'autre part

Très utilisée notamment en régression où l'on cherche à expliquer la variation d'une variable de sortie (expliquée) Y, par la variation d'une variable d'entrée (explicative, covariable) X

Compte tenu de la valeur de X, la meilleure prédiction de Y (en termes d'erreur quadratique) est l'espérance $f(X)$ de Y sachant X.

On dit que Y est une fonction de X plus un bruit (erreur) :

$$Y = f(X) + E \tag{1}$$

f est appelée la fonction de régression, et E est un bruit souvent supposé d'espérance nulle.

Méthode des Moindres Carrées II

L'estimateur des MC à des propriétés optimales d'absence de biais, de variance minimale (sous certaines conditions)

Critère des moindres carrés

Soit le modèle

$$Y_i = f(X_i) + E_i \tag{2}$$

La fonction f est à estimer à partir d'un échantillon des couples de covariables X_i et leur réponses Y_i : $((x_1, y_1), ..., (x_n, y_n))$

Cette estimation est effectuée en minimisant la somme des écarts (erreurs) quadratiques

Définition : Critères des moindres carrés

L'erreur quadratique est donnée par la somme des carrés des résidus (Residual Sum of Squares (RSS)) :

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 . \tag{3}$$

Moindres Carrés

Erreur quadratique dans le cas d'une fonction paramétrique

Soit $f(x;\theta)$ une fonction de paramètre θ à estimer. La somme des écarts qudratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^{n} \left(y_i - f(x_i;\theta)\right)^2 . \tag{4}$$

Moindres Carrés

Erreur quadratique dans le cas d'une fonction paramétrique

Soit $f(x;\theta)$ une fonction de paramètre θ à estimer. La somme des écarts qudratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 . \quad (4)$$

Définition : **Définition de l'estimateur des moindres carrés**

L'estimation de θ par la méthode des moindres carrés consiste à choisir, comme estimation de θ, la valeur de θ qui minimise la fonction $\mathrm{RSS}(\theta)$.

$$\hat{\theta} = \arg\min_{\theta} \mathrm{RSS}(\theta) \quad (5)$$

En effet, en choisissant une valeur de θ qui minimise $\mathrm{RSS}(\theta)$, cela revient à dire que, parmi les valeurs possible de θ, nous prenons la valeur qui correspond à une erreur minimale que les réponses y s'écartent de $f(x;\theta)$ pour l'échantillon observé $((x_1, y_1), \ldots, (x_n, y_n))$.

Moindres carrés : Cas d'un seul paramètre θ

Estimateur des moindres carrés (MC)

L'*estimateur de moindres carrés* (MC) de θ, noté $\hat{\theta}$, à partir des valeurs de l'échantillon $((x_1, y_1), \cdots, (x_n, y_n))$ peut être déterminé à partir de

$$\frac{\mathrm{d}\,\mathrm{RSS}(\theta)}{\mathrm{d}\,\theta}\Big|_{\theta=\hat{\theta}} = 0. \tag{6}$$

qui permet d'identifier les extrema de $\mathrm{RSS}(\theta)$. Il faut donc, après que les solutions de l'équation aient été trouvées, sélectionner celles qui correspondent à des minima. Un minimum vérifie

$$\frac{\mathrm{d}^2\,\mathrm{RSS}(\theta)}{\mathrm{d}^2\,\theta}\Big|_{\theta=\hat{\theta}} > 0 \tag{7}$$

il faut donc sélectionner parmi les solutions de la première équation celles qui vérifient cette deuxième équation.

Moindres Carrés

⚠ Remarque :

Bien que la plupart des critères d'EQ soient différentiables, la minimisation du critère des MC ne s'effectue pas toujours de façon analytique

⇒ On a souvent recours à des méthodes d'optimisations numériques (par exemple comme en réseau de neurones, etc)

⇒ la descente de gradient, l'algorithme de Newton Raphson, etc

Dans le cas où la fonction d'erreur est convexe, l'estimateur du Moindres Carrés fournit le minimum global. Cependant, dans beaucoup de problèmes réels, la fonction d'erreur n'est pas convexe et l'on a un minimum local ; atteindre le minimum global n'est pas toujours garanti

Des procédures algorithmiques existent (plusieurs initialisations, etc) et peuvent permettre d'atteindre un “bon” minimum local

Moindres Carrés : cas de paramètres multiples I

Dans le cas d'un paramètre multiple $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$,le critère d'erreur est donné par

$$\mathsf{RSS}(\boldsymbol{\theta}) = \mathsf{RSS}(\theta_1, \ldots, \theta_m)$$

Les estimateurs de MC de $\theta_j, j = 1, \cdots, m$, sont obtenus en résolvant simultanément le système d'équations suivant

$$\frac{\partial \mathsf{RSS}(\boldsymbol{\theta})}{\partial \theta_j}\Big|_{\theta_j = \hat{\theta}_j} = 0 \quad \text{pour } j = 1, \ldots, m \tag{8}$$

Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

Absence de biais

Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle ($\mathbb{E}[E_i] = 0$), l'estimateur des MC est sans bias

Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

Absence de biais

Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle ($\mathbb{E}[E_i] = 0$), l'estimateur des MC est sans bias

variance minimale

Si les erreurs sont d'espérance nulle ($\mathbb{E}[E_i] = 0$) et homoscédastiques décorrélées ($\mathbb{E}[E_i^T E_i] = \sigma^2 \mathbf{I}$) L'EMC est alors à variance minimale

$\Rightarrow$ efficace et est donc le meilleur estimateur sans biais

Ces propriétés sont valables quelle que soit la distribution des erreurs.

Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

Absence de biais

Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle ($\mathbb{E}[E_i] = 0$), l'estimateur des MC est sans bias

variance minimale

Si les erreurs sont d'espérance nulle ($\mathbb{E}[E_i] = 0$) et homoscédastiques décorrélées ($\mathbb{E}[E_i^T E_i] = \sigma^2 \mathbf{I}$) L'EMC est alors à variance minimale

$\Rightarrow$ efficace et est donc le meilleur estimateur sans biais

Ces propriétés sont valables quelle que soit la distribution des erreurs.

Si en plus on fait l'hypothèse de normalité sur les erreurs ($e_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$) :

Normalité

La distribution de $\hat{\Theta}$ est normale centrée sur le vrai paramètre θ

Régression linéaire

- Introduction
- Le modèle linéaire simple
- Estimation par moindres carrés
- Formulation vectorielle

Régression linéaire I

L'estimation de paramètres et l'évaluation de la qualité d'un estimateur seront appliqués dans se chapitre à une famille de problèmes très connues en statistique et estimation qui est celle de la régression simple et multiple.

Une situation qu'on rencontre couramment est celle dans laquelle une variable aléatoire Y est fonction d'une ou plusieurs variables indépendantes (déterministes) $(x_1, \ldots, x_m)$.

Par exemple le prix d'un logement (Y) est une fonction de sa localisation (x_1) et de son âge (x_2) ;

la durée de vie d'un composant électronique (Y) peut être liée à la température (x_1), la pression (x_2), etc ; la vitesse d'un automobiliste (Y) en fonction du temps t, etc.

Régression linéaire II

Notons que les variables indépendantes est aussi appelées **variables explicatives** car à travers elles on cherche à expliquer les variables Y qui sont dites **expliquées**[1]

L'objectif est donc d'estimer "la relation" entre Y et les variables indépendantes $(x_1, \ldots, x_m)$ étant donnés un échantillon des couples $((\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n))$ des variables

$(Y_1, \ldots, Y_n)$ de la variable Y et les valeurs associées des variables explicatives $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$, telle que x_j, $j = 1, \ldots, m$ pour chaque valeurs observée de Y_i, $i = 1 \ldots, n$.

1. En informatique et en particulier en machine learning (apprentissage), on trouve aussi l'appellation entrées/sorties pour respectivement variables explicatives et expliquées.

Le modèle linéaire simple I

prenons le cas simple où l'on suppose que Y ne dépend que d'une seule variable explicative x et que cette relation est supposée linéaire. en d'autres termes on a la relation suivante

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{9}$$

avec $(\beta_0, \beta_1) \in \mathbb{R}^2$ sont les deux paramètres de la droite de régression, β_0 appelé *ordonnée à l'origine (intercept)* et β_1 *pente (slope)*. Ce sont les **coefficients de régression**

ϵ est une variable aléatoire représentant un résidu (erreur de mesure).

En effet, ce modèle suppose que la variable expliquée que l'on observée résulte du vrai modèle (ici le modèle linéaire représentée par la droite) et un bruit (de mesure par exemple) ou tout autre type d'erreur.

Le modèle linéaire simple I

Le bruit est généralement supposé d'espérance nulle et de variance σ^2 et décorrélé ($\text{cov}(\epsilon_i, \epsilon_j)_{i \neq j} = 0$).

Dans ce cas σ^2 devient également un paramètre du modèle et est donc aussi à estimer.

Dans le cadre de ce cours, on va supposer que ce bruit est en plus Gaussien. Il en découle donc qu'il est indépendant (les ϵ_i dont i.i.d).

Les deux paramètres (β_0, β_1) sont inconnus et donc à estimer.

Cette estimation sera effectuée à partir d'un échantillon de couples $((x_1, Y_1), \ldots, (x_n, Y_n))$ [2].

2. Ici nous utilisons la notations (x_i, Y_i) vu que x est déterministe mais cela ne change rien au modèle si X est aléatoire.

Le modèle linéaire simple I

Le modèle s'écrit donc sous la forme

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{10}$$

où ϵ_i est le bruit associée à la *i*ème variable aléatoire. Étant donnés donc une réalisation (valeur) y_i de chaque Y_i et le résidu associé (réalisation de la variable aléatoire représentant le bruit) que nous notons e_i on obtient alors :

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{11}$$

Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés

mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle (par exemple par maximum de vraisemblance ou encore par inférence bayésienne (maximum a posteriori)).

Le modèle linéaire simple : Estimation par moindres carrés

La méthode des moindres carrés est une approche d'estimation ponctuelle des paramètres de régression (β_0, β_1).

fournit les estimations $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimisent la somme des écarts quadratiques entre les valeurs observées y_i et l'espérance $\beta_0 + \beta_1 x_i$ du modèle de Y_i.

D'après (11), l'écart entre la valeur d'une observation et l'espérance du modèle est

$$e_i = y_i - (\beta_0 + \beta_1 x_i).$$

La somme des carrés des résidus est donc donnée par

$$\mathrm{RSS}(\beta_0, \beta_1) = Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2. \tag{12}$$

$\mathrm{RSS}(\beta_0, \beta_1)$ est quadratique $\Rightarrow$ minimisation analytique

Le modèle linéaire simple : Estimation par moindres carrés

Theorem

Considérons le modèle de régression simple donné par l'équation (9). Soit $((x_1, y_1), \ldots, (x_n, y_n))$ un échantillon de valeurs observées de Y et leurs veleurs associées de x. Alors les estimations des moindres carrées ordinaires (MCO) de β_0 et β_1 sont données par :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{13}$$

$$\hat{\beta}_1 = \left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right] \left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{-1}, \tag{14}$$

où $\bar{x}$ représente la moyenne empirique des x_i et $\bar{y}$ la moyenne moyenne empirique es y_i et sont donnés par $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Le modèle linéaire simple : Estimation par moindres carrés

Preuve. Selon les MCO, les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont celles qui minimisent la somme des carrées des résidus (12) par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$.
Mathématiquement on note

$$\begin{aligned}
(\hat{\beta}_0, \hat{\beta}_1) &= \arg \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \mathrm{RSS}(\beta_0, \beta_1) \\
&= \arg \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 . \qquad (15)
\end{aligned}$$

Nous aurons donc déterminé les estimations $(\hat{\beta}_0, \hat{\beta}_1)$. Ainsi, nous avons :

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_0} = -2 \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big) \\
\frac{\partial Q}{\partial \beta_1} &= \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \big(y_i - (\beta_0 + \beta_1 x_i)\big) \cdot
\end{aligned}$$

Le modèle linéaire simple : Estimation par moindres carrés

Ensuite, en annulant ces dérivés nous avons :

$$\begin{aligned}
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i &= 0 \\
\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= 0 \cdot
\end{aligned}$$

ce qui donne les **équations normales**

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \tag{16}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{17}$$

Le modèle linéaire simple : Estimation par moindres carrés

1. En divisant par n la première équation on obtient la valeur de $\hat{\beta}_0$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \bar{y} - \hat{\beta}_1 \bar{x} \,.$$

2. En remplaçant $\hat{\beta}_0$ par sa valeur on obtient

$$\sum_{i=1}^{n} x_i \bar{y} - \hat{\beta}_1 \sum_{i=1}^{n} x_i \bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \,,$$

ce qui donne enfin

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \,.$$

Le modèle linéaire simple : Estimation par moindres carrés

Il reste à vérifier si la dérivée partielle seconde de Q par rapport à au moins l'un des paramètres est positive et que le déterminant de la matrice des dérivées partielles secondes de Q est strictement positif.

- On a

$$\begin{aligned}\frac{\partial^2 Q}{\partial^2 \beta_0} &= \frac{-2\partial \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)}{\partial \beta_0} \\ &= -2\sum_{i=1}^{n} \frac{\partial \beta_0}{\partial \beta_0} = 2n \geq 0\cdot\end{aligned}$$

- Le déterminant de la matrice des dérivées partielles secondes de Q est :

$$\det \begin{pmatrix} \frac{\partial^2 Q}{\partial^2 \beta_0} & \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 Q}{\partial^2 \beta_1} \end{pmatrix} = \det \begin{pmatrix} 2n & 2\sum_i x_i \\ 2\sum_i x_i & 2\sum_i x_i^2 \end{pmatrix} = 4n\sum_i (x_i - \bar{x})^2 > 0\cdot$$

Notez que ce déterminant est nul si tous les x_i prennent la même valeur.
$\Rightarrow$Au moins deux valeurs x_i distinctes sont nécessaires pour la détermination de (β_0, β_1).

Régression linéaire : Formulation vectorielle

maintenant nous reformulons le modèle que nous venons de voir sous forme de vecteurs-matrices.

Comme nous allons le voir, les résultats sous la forme matricielle sont obtenus à partir de calculs simples.

Cela permettra aussi de généraliser le modèle linéaire simple à des modèles généraux notamment la régression multiple.

Soit $(y_1, \ldots, y_n)$ l'ensemble des valeurs observées de la variable dépendante Y et $(x_1, \ldots, x_n)$ l'ensemble des valeurs observées de la variable explicative x.

⚠ Remarque : L'ensemble des couples $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$ s'appelle aussi *ensemble d'apprentissage*. Car c'est l'ensemble de données à partir duquel on va estimer notre modèle (donc *apprendre le modèle*) pour pouvoir ensuite prédire la valeur de Y pour une nouvelle valeur de x

Régression linéaire : Formulation vectorielle

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + e_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + e_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + e_n \ \cdot \end{aligned} \tag{18}$$

Régression linéaire : Formulation vectorielle

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + e_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + e_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + e_n \ \cdot \end{aligned} \tag{18}$$

Notons par

- $\mathbf{y} = (y_1, \ldots, y_n)^T$ le vecteur des valeurs d'observations de Y,
- $\mathbf{e} = (e_1, \ldots, e_n)^T$ le vecteur des valeurs des résidus
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ le vecteur des coefficients de régressions à estimer
- $\mathbf{X}$ la *matrice de régression* (matrice de *design* ou de *Vendermonde*)

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

Régression linéaire : Formulation vectorielle

le modèle (18) s'écrit donc sous la forme matricielle suivante :

Régression linéaire : Formulation vectorielle

$$y = \mathbf{X}\beta + \mathbf{e} \cdot \tag{19}$$

La somme des écarts quadratiques $\sum_{i=1}^{n} e_i^2$ est maintenant donnée par :

Régression linéaire : Formulation vectorielle

$$\text{RSS}(\beta) = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{20}$$

Régression linéaire : Formulation vectorielle

L'estimation $\hat{\boldsymbol{\beta}}$ par moindres carrés de $\boldsymbol{\beta}$ s'obtient en minimisant (20) qui est une fonction quadratique en $\boldsymbol{\beta}$.

On a :

- $\text{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
- $\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
- $\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \mathbf{0} \Rightarrow -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$

$\Rightarrow$ les équations normales

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}\,. \tag{21}$$

Régression linéaire : Formulation vectorielle

$$
\begin{aligned}
\mathbf{X}^T\mathbf{X}\hat{\beta} &= \mathbf{X}^T\mathbf{y} \\
(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
$$

$\Rightarrow$ Estimateur des moindres carrés :

$$
\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{22}
$$

Notez que l'inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ existe si les données comportent au moins deux valeurs distinctes de x_i.