Régression multiple

Introduction

- Étudier la liaison entre une "variable à expliquer" quantitative Y et une suite de "variables explicatives" quantitatives $X_1 \dots X_p$
- Une variable quantitative Y est modélisée par plusieurs variables quantitatives X_j $(j = 1, \dots, p)$

Modèle

- Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n ($n > p + 1$) de $R^{(p+1)}$:
  - (x_{1i} , . . . , x_{ji} , . . . , x_{pi} , y_i) i = 1, …, n.
  - $Y = b_0 + b_1X_1 + \ldots + b_pX_p + \varepsilon$
  - $y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_px_{pi} + \varepsilon_i$
  - b_j = coefficients de régression fixes, sont supposés constants (mais inconnus)
  - ε = sont des termes d'erreur, non observés, indépendants et identiquement distribués, ils sont aléatoires N ($0, \sigma^2$). :
    - de moyenne 0
    - d'écart - type σ

Vocabulaire

- Y
  - Variable à expliquer
  - Variable dépendante
  - Variable endogène
- X_1 X_2 ... X_p
  - Variables explicatives
  - Variables indépendantes
  - Variables exogènes

Représentation matricielle des données

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

- x_{ij} représente la valeur prise par la variable explicative j sur l'unité statistique i.
- Le vecteur $y = (y_1 \, . \, . \, . \, y_i \, . \, . \, . \, y_n)'$ représente les valeurs prises par la variable dépendante sur les n unités statistiques.

Représentation matricielle des données

- Dans la plupart des applications, on supposera également que la première variable est la constante,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Régression multiple: exemples

Expliquer	en fonction de
Prix d'un appartement	Superficie standing quartier sécurité proximité de commerce
Prix d'une voiture	cylindrée taille vitesse maximum origine niveau de finition
Prévoir des ventes	budget de recherche investissements publicité remises aux grossistes prix de vente

Estimation des paramètres du modèle: le modèle

- L'écriture du modèle linéaire conduit à supposer que l'espérance de Y appartient au sous-espace de R^n engendré par $\{1, X_1, \ldots, X_p\}$ où 1 désigne le vecteur de R^n constitué de "1".
- Nous considérons donc que y_i est la réalisation d'une variable aléatoire Y_i définie par :
  - $Y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + \varepsilon_i$

Estimation des paramètres: b_0 b_1... b_p

- Nous recherchons les estimations : $\hat{b}_0 \hat{b}_1 \ldots \hat{b}_P$ des paramètres b_0 b_1 … b_p permettant de reconstituer au mieux les données y_i à partir des observations des p variables X_1 … X_p pour l'individu i

Éléments de solution

- La résolution du problème passe par la résolution d'un système de (p + 1) équations à inconnues
- Elle peut se présenter sous forme matricielle en utilisant les notations matricielles ci-dessous:

$$Y' = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{(n,\, p+1)}{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \\ 1 & x_{i1} & \dots & x_{ip} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

$$b' = \begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_p \end{bmatrix} \qquad \varepsilon' = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Principe des moindres carrés

- La régression de y en X consiste à chercher un vecteur de coefficients de régression :

  $b = (b_1, \dots, b_j, \dots, b_p)$
- qui permet d'ajuster au mieux les y_i par les x_{ij} au sens où la somme des carrées des résidus est minimisée. La régression s'écrit :

$$
\begin{aligned}
y_i &= b_1 + b_2 x_{i2} + \cdots + b_j x_{ij} + \cdots + b_p x_{ip} + e_i \\
&= \sum_{j=1}^{p} b_j x_{ij} + e_i, i = 1, \ldots, n.
\end{aligned}
$$

- Cette équation peut s'écrire de manière matricilelle :

  $y = Xb + e.$

Principe des moindres carrés

- La somme des carrés des résidus est le critère à minimiser

$$
\begin{aligned}
Q(b_1, \ldots, b_p) &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n} \left[y_i - (b_1 + b_2 x_{i2} + \cdots + b_j x_{ij} + \cdots + b_p x_{ip})\right]^2 \\
&= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} b_j x_{ij} \right)^2 .
\end{aligned}
$$

Principe des moindres carrés

- Il est cependant plus facile d'utiliser une écriture matricielle. Comme on peut écrire

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = \begin{pmatrix} y_1 - \sum_{j=1}^{p} b_j x_{1j} \\ \vdots \\ y_i - \sum_{j=1}^{p} b_j x_{ij} \\ \vdots \\ y_n - \sum_{j=1}^{p} b_j x_{nj} \end{pmatrix}$$

Moindres carrés : solution

- la régression de y en X au sens des moindres carrés consiste à chercher l'ajustement qui minimise en b :

$$\begin{aligned} Q(b_1,\dots,b_p) = Q(\mathbf{b}) &= \sum_{i=1}^{n} e_i^2 \\ &= \mathbf{e}'\mathbf{e} = (\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) \\ &= ||\mathbf{y}-\mathbf{X}\mathbf{b}||^2. \end{aligned}$$

- Remarquons que Q(b) peut également s'écrire

$$\begin{aligned} Q(\mathbf{b}) &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}, \end{aligned}$$

- car y 'Xb = b 'X'y

Moindres carrés : solution

- Le vecteur de coefficient de régression au sens des moindres carrés est donné par

$$b = (X'X)^{-1}\,X'y,$$

- sous réserve que la matrice $X'X$ soit inversible.

Démonstration

- Calculons le vecteur des dérivées partielles Q(b) par rapport à b,

$$\frac{\partial Q(b)}{\partial b} = \frac{\partial y' y}{\partial b} - \frac{\partial 2 y' Xb}{\partial b} + \frac{\partial b' X' Xb}{\partial b}$$

$$= 0 - 2X' y + 2X' Xb = 2X' Xb - 2X' y$$

- En annulant le vecteur des dérivées partielles, on obtient un système de p équations à p inconnues

$$\frac{\partial Q(b)}{\partial b} = 2X' Xb - 2X' y = 0$$

Démonstration

- ce qui donne : $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}.$
- En faisant l'hypothèse que X'X est inversible, on peut déterminer b :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- Pour montrer que l'on a bien obtenu un minimum, on calcule la matrice hessienne des dérivées secondes

$$\mathbf{H} = \frac{\partial^2 Q(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X}.$$

- Cette matrice est définie positive, on a donc bien un minimum.

Valeurs ajustées et résidus

- Le vecteur des valeurs ajustées est le vecteur des prédictions ou estimations de y au moyen de X et de b, c'est-à-dire

$$\mathbf{y}^* = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- Le vecteur des valeurs ajustées peut être interprété comme la projection de y sur le sous-espace engendré par les colonnes de la matrice X.

Valeurs ajustées et résidus

- Le vecteur des résidus est la différence entre y et y* .

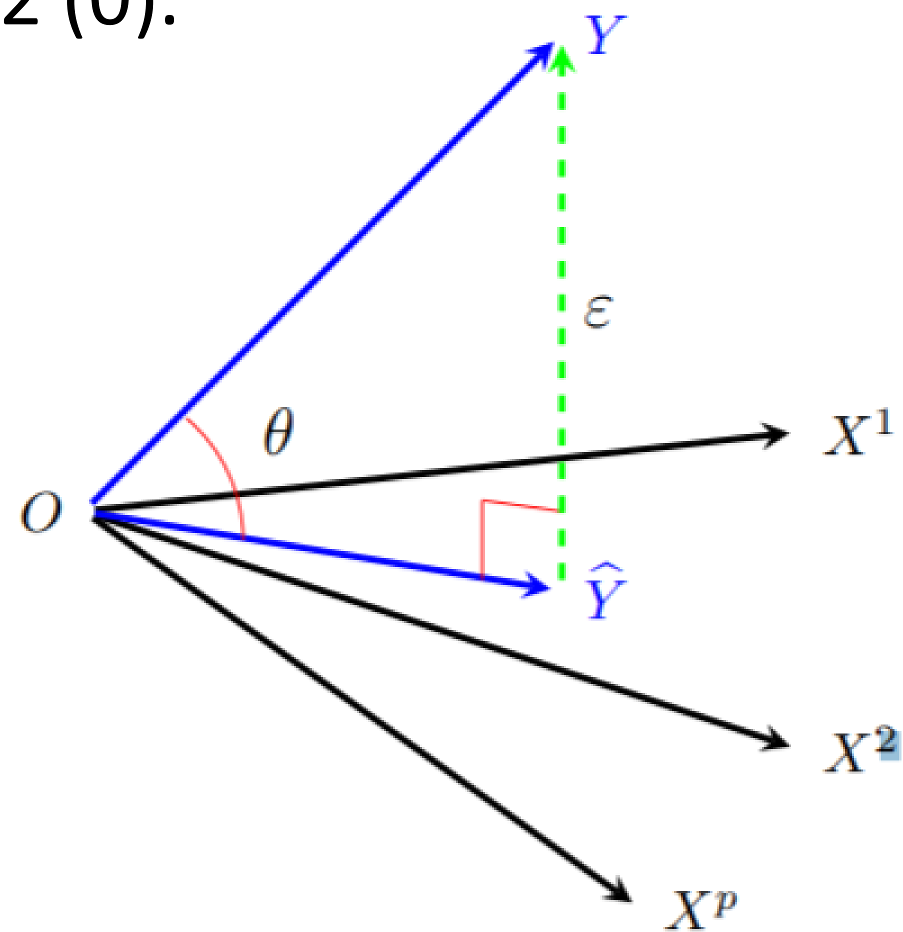
$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{y}^* = \mathbf{y} - \mathbf{X}\mathbf{b}\\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}.\end{aligned}$$

Valeurs ajustées et résidus

- Propriété
  - $y = y^* + e$,
  - y^* est une combinaison linéaire des colonnes de X,
  - y^* et e sont orthogonaux,
  - e est orthogonal avec toutes les colonnes de X, c'est-à-dire $e'X = 0$.

Géométrie

- La régression est la projection $\hat{Y}$ de Y sur l'espace vectoriel Vect{1, X_1 , . . . , X_p} ; de plus R^2 = cos2 (θ).



Variance de régression et variance résiduelle

- Soit le vecteur de R^n contenant n fois la moyenne de la variable y : $$\bar{\mathbf{y}} = (\bar{y}, \ldots, \bar{y})'.$$
- La variance peut être définie simplement par :

$$s_y^2 = \frac{1}{n}(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

- La variance de régression est la variance des valeurs ajustées :

$$s_Y^2 = \frac{1}{n}(\mathbf{y}^* - \bar{\mathbf{y}})'(\mathbf{y}^* - \bar{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \bar{y})^2 .$$

- La variance résiduelle est la variance des résidus :

$$s_e^2 = \frac{1}{n}\mathbf{e}'\mathbf{e} = \frac{1}{n}(\mathbf{y} - \mathbf{y}^*)'(\mathbf{y} - \mathbf{y}^*) = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

Coefficient de détermination

- Le coefficient de détermination vaut

$$R^2 = \frac{s_Y^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

- Il est important de noter que le R^2 ne peut être calculé que si la régression inclut une constante. Si ce n'est pas le cas, le R^2 peut prendre des valeurs négatives.
- La racine carrée du coefficient de détermination est appelée le coefficient de corrélation multiple entre la variable à expliquer Y et les variables explicatives $X_1, \dots X_p$.

Propriétés

$$R^2 = \frac{\text{S. C. expliquée}}{\text{S. C. totale}}$$

$$= 1 - \frac{\text{S. C. résiduelle}}{\text{S. C. totale}}$$

$$0 \leq R^2 \leq 1$$

R^2 proche de 1 signifie :

Y est "bien expliquée" par les variables $X_1 \ldots X_k$

La précision des résultats

- On montre que cette précision est liée à celle de s_Y^2 .
- Plus s_Y^2 est faible, plus les coefficients ont des chances d'être estimés précisément.
- A chaque estimation d'un coefficient, est associé un écart-type estimé à partir des données traitées (en général noté erreur-type).
- On peut comparer ces erreurs-types aux coefficients estimés.

EXEMPLES

Exemple 1: cigarette

constante	TAR (mg)	NICOTINE (mg	WEIGHT (g)	CO (mg)
1	14.1	0.86	0.9853	13.6
1	16	1.06	1.0938	16.6
1	8	0.67	0.928	10.2
1	4.1	0.4	0.9462	5.4
1	15	1.04	0.8885	15
1	8.8	0.76	1.0267	9
1	12.4	0.95	0.9225	12.3
1	16.6	1.12	0.9372	16.3
1	14.9	1.02	0.8858	15.4
1	13.7	1.01	0.9643	13
1	15.1	0.9	0.9316	14.4
1	7.8	0.57	0.9705	10
1	11.4	0.78	1.124	10.2
1	9	0.74	0.8517	9.5
1	1	0.13	0.7851	1.5
1	17	1.26	0.9186	18.5
1	12.8	1.08	1.0395	12.6
1	15.8	0.96	0.9573	17.5
1	4.5	0.42	0.9106	4.9
1	14.5	1.01	1.007	15.9
1	7.3	0.61	0.9806	8.5
1	8.6	0.69	0.9693	10.6
1	15.2	1.02	0.9496	13.9
1	12	0.82	1.1184	14.9

- X'X

24	275.6	19.88	23.0921
275.6	3613.16	254.177	267.46174
19.88	254.177	18.0896	19.266811
23.0921	267.46174	19.266811	22.3637325

- $(X'X)^{-1}$

6.56299	0.06290	-0.93908	-6.71991
0.06290	0.02841	-0.45200	-0.01528
-0.93908	-0.45200	7.86328	-0.39900
-6.71991	-0.01528	-0.39900	7.50993

- X'Y

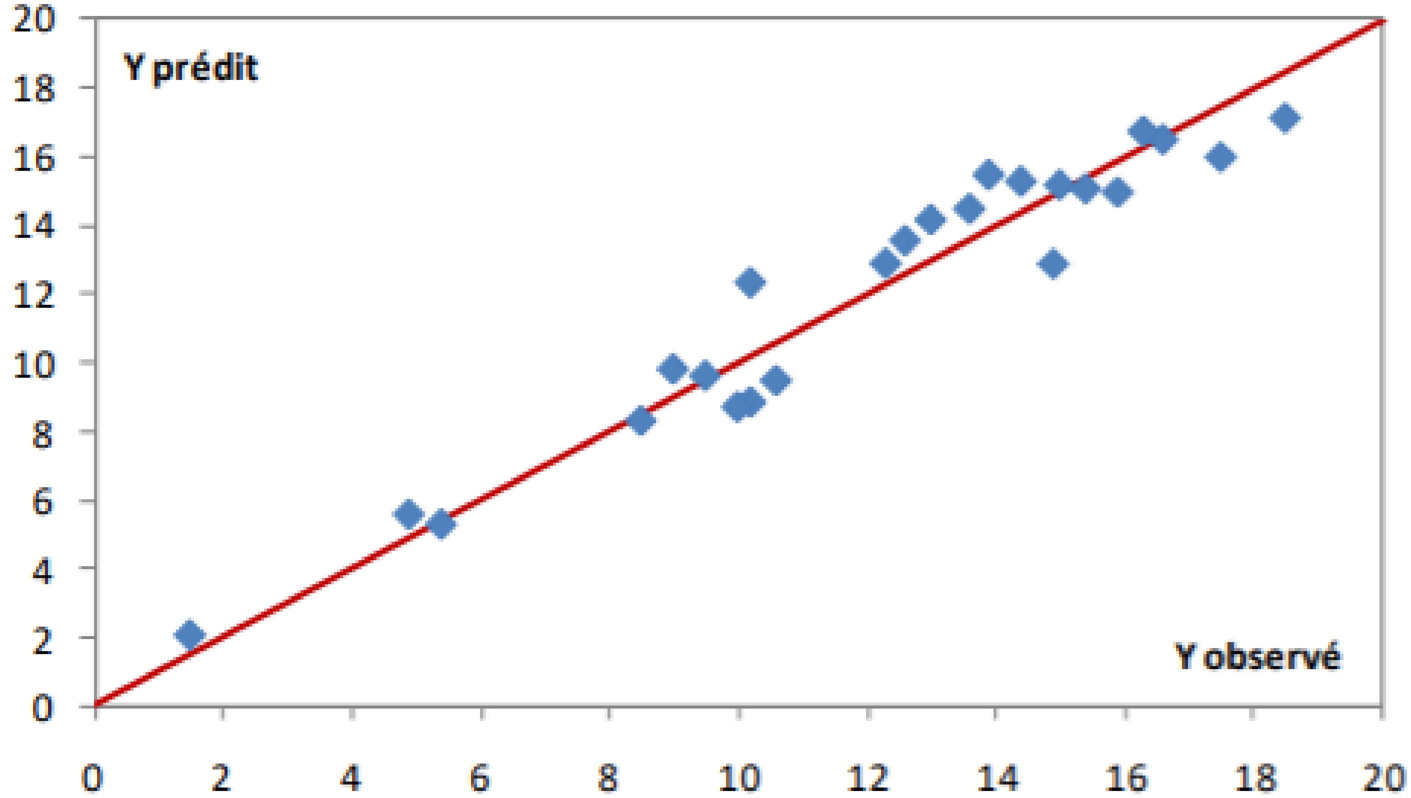
289.7
3742.85
264.076
281.14508

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

-0.55170	constante
0.88758	tar
0.51847	nicotine
2.07934	weight

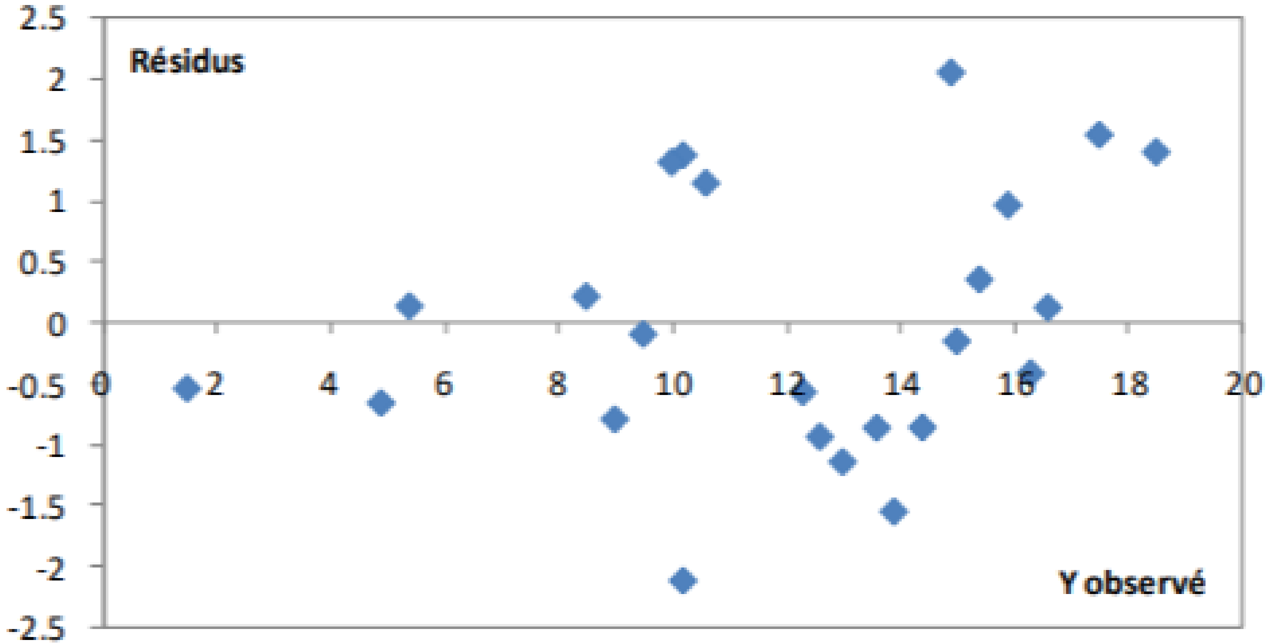
Diagnostic graphique

- Y observé vs. Y prédit



Diagnostic graphique

- Y observé vs. résidu



Exemple 2: Hauteur de neige

- Dans l’exemple qui suit nous réalisons une régression multiple pour expliquer la hauteur de neige en fonction de l’altitude, de la rugosité, de la pente, de l’orientation, de la latitude et de la longitude

H_NEIGE	vecteur	altitude	rugosite	pente	orient.	lat	long.
95	1	2768	252	22	324	8760219	438465.0625
150	1	4108	333	29	308	8760195	438474.0625
4	1	4045	62	5	249	8760168	438480.0625
0	1	4572	85	8	14	8760135	438489.0625
0	1	4614	115	10	63	8760105	438495.0625
80	1	4321	176	16	130	8760072	438498.0625
95	1	3886	72	6	199	8760039	438504.0625
20	1	4206	57	5	32	8760012	438507.0625
90	1	4192	266	23	197	8759985	438513.0625
10	1	4051	69	6	113	8759955	438519.0625
10	1	3746	62	5	149	8759922	438519.0625
50	1	3789	42	3	218	8759895	438525.0625
45	1	3771	44	4	53	8759865	438531.0625
60	1	3796	48	4	101	8759838	438534.0625
55	1	3885	77	7	332	8759811	438537.0625
3	1	4295	113	10	18	8759787	438540.0625
33	1	4467	147	13	50	8759760	438546.0625
0	1	4764	12	1	276	8759730	438552.0625
35	1	4313	38	3	350	8759703	438552.0625
45	1	4387	40	3	46	8759673	438558.0625

- Le produit X'X:

20.0000	81976.0000	2110.0000	183.0000	3222.0000	175198869.0000	8770339.2500
81976.0000	339594498.0000	8487334.0000	736618.0000	12861325.0000	718104679425.0000	35947950323.5000
2110.0000	8487334.0000	366956.0000	32036.0000	386290.0000	18483638688.0000	925244282.8750
183.0000	736618.0000	32036.0000	2799.0000	33323.0000	1603083666.0000	80246258.4375
3222.0000	12861325.0000	386290.0000	33323.0000	771684.0000	28224580695.0000	1412891754.3750
175198869.0000	718104679425.0000	18483638688.0000	1603083666.0000	28224580695.0000	1534732185500860.0000	76827675778567.3000
8770339.2500	35947950323.5000	925244282.8750	80246258.4375	1412891754.3750	76827675778567.3000	3845942542298.3300

- $(X'X)^{-1}$:

42548515331.8374	73.5283	-569.7835	4096.6641	-164.4807	-3668.8247	-23739.2652
73.5284	0.0000	0.0000	-0.0001	0.0000	0.0000	0.0000
-569.7830	0.0000	0.0047	-0.0535	0.0000	0.0001	0.0003
4096.6572	-0.0001	-0.0535	0.6061	0.0005	-0.0004	-0.0014
-164.4807	0.0000	0.0000	0.0005	0.0000	0.0000	0.0001
-3668.8247	0.0000	0.0001	-0.0004	0.0000	0.0003	0.0020
-23739.2657	0.0000	0.0003	-0.0014	0.0001	0.0020	0.0133

- $X'y$:

880
3458806
140963
12244
181900
7708792743
385887448

- Donc $(X'X)^{-1}X'y$ donne les termes de l'équation multiple :
  - Constante : -6111180.498
  - Altitude : -0.03526
  - Rugosité : 1.0379
  - Pente : -7.6228
  - Orientation : 0.0907
  - Latitude : 0.5191
  - Longitude : 3.6401

Exemple 3: Production, travail et capital

- Les données présentées dans le tableau ci-dessous concernent 9 entreprises de l'industrie chimique. Nous cherchons à établir une relation entre la production y_i, les heures de travail x_{i1} et le capital utilisé x_{i2}.

Entreprise i	Travail (heures) x_{i1}	Capital (machines/heures) x_{i2}	Production (100 tonnes) y_i
1	1 100	300	60
2	1 200	400	120
3	1 430	420	190
4	1 500	400	250
5	1 520	510	300
6	1 620	590	360
7	1 800	600	380
8	1 820	630	430
9	1 800	610	440

- Le modèle de régression linéaire multiple avec 2 variables explicatives est donc :

$$y = b_0 + b_1x_1 + b_2x_2 + \varepsilon$$

- La notation matricielle :

$$y = X\beta + \varepsilon$$

$$\mathbf{y} = \begin{bmatrix} 60 \\ 120 \\ 190 \\ 250 \\ 300 \\ 360 \\ 380 \\ 430 \\ 440 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1\,100 & 300 \\ 1 & 1\,200 & 400 \\ 1 & 1\,430 & 420 \\ 1 & 1\,500 & 400 \\ 1 & 1\,520 & 510 \\ 1 & 1\,620 & 590 \\ 1 & 1\,800 & 600 \\ 1 & 1\,820 & 630 \\ 1 & 1\,800 & 610 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{bmatrix}$$

- Nous allons calculer le vecteur des estimateurs β_b défini par l'égalité suivante :

$$\beta_b = ({}^t XX)^{-1\,t}Xy.$$

- Donc

$$({}^t\mathbf{X}\mathbf{X}) = \begin{bmatrix} 9 & 13\,790 & 4\,460 \\ 13\,790 & 21\,672\,100 & 7\,066\,200 \\ 4\,460 & 7\,066\,200 & 2\,323\,600 \end{bmatrix}$$

$$({}^t\mathbf{X}\mathbf{X})^{-1} = \begin{bmatrix} 6,304\,777 & -0,007\,800 & 0,011\,620 \\ -0,007\,800 & 0,000\,015 & -0,000\,031 \\ 0,011\,620 & -0,000\,031 & 0,000\,072 \end{bmatrix}$$

- Et

$${}^t\mathbf{X}\mathbf{y} = \begin{bmatrix} 2\,530 \\ 4\,154\,500 \\ 1\,378\,500 \end{bmatrix}$$

- Ainsi

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = ({}^{\mathbf{t}}\mathbf{X}\mathbf{X})^{-1\,\mathbf{t}}\mathbf{X}\mathbf{y} = \begin{bmatrix} -437,714 \\ 0,336 \\ 0,410 \end{bmatrix}$$

- L’équation des moindres carrés de notre modèle est donc:

$$\hat{y}(x_1, x_2) = -437,714 + 0,336\,x_1 + 0,410\,x_2$$