

Statistiques

Année 3 Semestre 6

Chapitre 3 : Régression Linéaire :
Simple (RLS) & Multiple (RLM)
au sens des Moindres Carrés (MC)

Laetitia DELLA MAESTRA

Enseignant-chercheur en Mathématiques

laetitia.dellamaestra@devinci.fr - Bureau L405

Références & Bibliographie

- Références données aux CMO précédents ;
- Equipe de Statistique de l'IRMAR (Institut de Recherche Mathématique de Rennes) ; attention : avec le logiciel R, mais présentation des méthodes & explications mathématiques de très grande qualité
Site de François Husson : <https://husson.github.io>

Livre Régression avec R, 2ème édition, P.-A. Cornillon, N. Hengartner, E. Matzner-Lober, L. Rouvière, ed. EDP Sciences
- Machine-Learning avec Scikit-Learn, 3ème édition, Aurélien Géron, ed. Dunod
- Exploration de données & méthodes stat., L. Bellanger et R. Tomassone, ed. Ellipses

Attention : les seules notations, définitions, appellations, etc... qui font foi dans ce module sont celles du présent cours.

Plan du cours

A) Présentation & Résolution (algébrique/géométrique & analytique)

Objectif 1 : **approximer**, avec les meilleurs paramètres possibles, le nuage de points représentant nos données quantitatives par une **forme affine**

ds \mathbb{R}^2 par une droite affine, ds \mathbb{R}^3 par un plan affine, ds \mathbb{R}^{p+1} ($p > 2$) par un hyperplan affine

Objectif 2 : effectuer des **prévisions** à l'aide du modèle approximé

Objectif 3 : évaluer la **qualité du modèle** choisi

Objectif 4 : **quantifier l'erreur** commise ds le choix de ce modèle approx

B) Modèle **probabiliste** des erreurs (on impose que l'erreur entre notre modèle théorique et nos données est aléatoire, les paramètres optimaux sont alors eux-mêmes aléatoires et sont des estimateurs des paramètres théoriques) :

Estimation des paramètres, calcul de leurs Espérance & Variance

↪ contrôle local d'ordre 1 et 2 de l'erreur commise avec le modèle choisi

C) Modèle **gaussien** des erreurs (on impose que l'erreur entre notre modèle théorique et nos données est gaussienne, les paramètres optimaux sont alors eux-mêmes gaussiens) :

Loi des paramètres, IC & Tests

↪ contrôle global de l'erreur commise avec le modèle choisi

D) **Limites du modèle** : Retour critique sur les conditions imposées

Objectif 5 : Sélection de variables ↪ **parcimonie** : peut-on faire quasiment aussi bien avec moins de facteurs explicatifs, c-à-d en dim + petite ?

A) Présentation

Problématique de la Régression

On dispose de données **quantitatives**, discrètes ou continues,

- concernant n **individus** $\mathcal{I}_1, \dots, \mathcal{I}_n$
- et $p + 1$ caractéristiques appelées **variables** x_1, \dots, x_p, y
 - y joue le rôle de la **variable à expliquer**
 - x_1, \dots, x_p jouent le rôle de **variables explicatives**

autrement dit on dispose de n séries de $p + 1$ valeurs :

$$(x_{1.} = (x_{11}, \dots, x_{1p}), y_1), \dots, (x_{n.} = (x_{n1}, \dots, x_{np}), y_n)$$

et l'on cherche à expliquer et à prédire y par x_1, \dots, x_p

- ⇒ on cherche quel type de relation il existe entre les x_i et les y_i
c-à-d qu'on cherche une fonction f telle que pour tout $i \in \llbracket 1, n \rrbracket$, $f(x_{i.}) \simeq y_i$
⇒ on cherche à prévoir le + précisément possible l'étiquette y_{n+1}^{prev} associée à un nouveau vecteur non étiqueté $x_{(n+1).}$ en posant $y_{n+1}^{\text{prev}} := f(x_{(n+1).})$.

Problématique de la Régression Linéaire

On cherche une **relation de type linéaire entre** x_i **et** y_i , càd que l'on se restreint aux fcts affines

$f : \mathbf{x}_i = (x_1, \dots, x_p) \in \mathbb{R}^p \mapsto \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ où $\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}$

autrement dit, l'on suppose que les points $(\mathbf{x}_i = (x_{i1}, \dots, x_{ip}), y_i)$ de l'espace affine \mathbb{R}^{p+1} sont regroupés autour d'un **hyperplan affine**

théorique d'équation $y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

Remarque : y_i peut être vue comme l'étiquette associée au i -ème vecteur \mathbf{x}_i : la Régression est une méthode d'Apprentissage

Supervisé

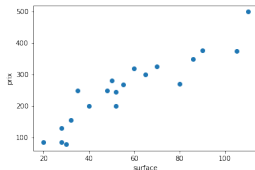
Exemple pour $p = 1$: Un agent immobilier reçoit un nouveau client lui demandant de prendre en charge la vente de son appartement. Pour fixer au mieux le prix de vente de ce bien en fonction de sa surface, l'agent étudie d'abord l'historique de ses ventes dans le quartier où est situé cet appartement : il s'agit de $n = 20$ appartements pour lesquels il dispose

- du prix \leadsto variable y
- de la surface \leadsto variable x_1 , que l'on notera x (il n'y a pas d'autre variable explicative !)

Objectif : expliquer y à partir de x

| Numéro i du bien | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Surface x_i (en m^2) | 28 | 50 | 55 | 110 | 60 | 48 | 90 | 35 | 86 | 65 |
| Prix de vente y_i (en K euros) | 130 | 280 | 268 | 500 | 320 | 250 | 378 | 250 | 350 | 300 |

| Numéro i du bien | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----------------------------------|-----|-----|-----|-----|----|----|-----|-----|-----|----|
| Surface x_i (en m^2) | 32 | 52 | 40 | 70 | 28 | 30 | 105 | 52 | 80 | 20 |
| Prix de vente y_i (en K euros) | 155 | 245 | 200 | 325 | 85 | 78 | 375 | 200 | 270 | 85 |



Autres exemples pour $p = 1$:

Exemple 2 pour $p = 1$: Une étude a été menée auprès de 12 étudiants afin d'expliquer le score à un examen d'anglais à partir du temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose du temps de révision en heures (variable x), du score obtenu sur 800 points (variable y)

| étudiant $n^{\circ} i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| x_i | 4 | 9 | 10 | 14 | 4 | 7 | 12 | 1 | 3 | 8 | 11 | 5 |
| y_i | 390 | 580 | 650 | 730 | 410 | 530 | 600 | 350 | 400 | 590 | 640 | 450 |

Ainsi avec une préparation de 4 heures, l'étudiant $n^{\circ} 1$ a obtenu le score de 390 à l'examen, avec une préparation de 9 heures, l'étudiant $n^{\circ} 2$ a obtenu le score de 580 à l'examen.

Exemple 3 pour $p = 1$: On étudie l'évolution du nombre d'inscriptions à un jeu en ligne au cours du temps. Pour chaque mois de l'année 2022, on dispose du rang du mois (variable x ; janvier est de rang 1, février est de rang 2, ...) et du nombre d'inscriptions en milliers (variable y). Les résultats sont

| $x_i = \text{mois } n^{\circ} i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|----------------------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| y_i | 37 | 43 | 41 | 40 | 51 | 47 | 48 | 54 | 56 | 64 | 66 | 73 |

Ainsi, au mois de janvier 2022, il y a eu 37000 inscriptions au jeu, en février 2022 il y a eu 43000 inscriptions au jeu.

Exemple 4 pour $p = 1$: Avant la commercialisation d'un produit, une entreprise effectue une étude de marché afin de déterminer la quantité demandée en milliers (variable y) en fonction du prix de vente en euros (variable x). Pour 6 prix de vente différents, les résultats sont :

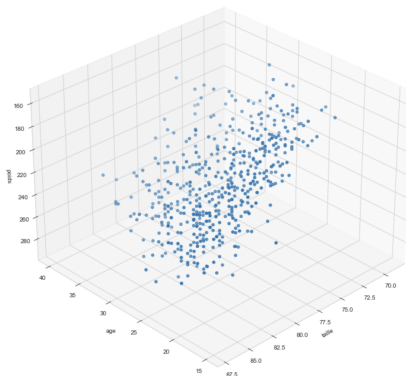
| $n^{\circ} i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---------------|------|------|------|------|-----|-----|
| x_i | 15 | 20 | 25 | 30 | 35 | 40 |
| y_i | 44.4 | 27.0 | 16.3 | 10.0 | 6.2 | 3.5 |

Exemple pour $p = 2$: On souhaite expliquer le poids d'un basketteur professionnel de la NBA à partir de sa taille et de son âge. Ainsi, pour $n = 505$ basketteurs de la NBA (données d'il y a dix ans), on dispose :

- de leur poids (en livres) \leadsto variable y
- de leur taille (en inches) \leadsto variable x_1
- de leur âge \leadsto variable x_2

Objectif : expliquer y à partir de x_1 et x_2

| | taille | age | poids |
|----------------|--------|-----|-------|
| Joueur | | | |
| Nate Robinson | 69 | 29 | 180 |
| Isaiah Thomas | 69 | 24 | 185 |
| Phil Pressey | 71 | 22 | 175 |
| Shane Larkin | 71 | 20 | 176 |
| Ty Lawson | 71 | 25 | 195 |
| ... | ... | ... | ... |
| Meyers Leonard | 85 | 21 | 250 |
| Rudy Gobert | 85 | 21 | 235 |
| Alex Len | 85 | 20 | 255 |
| Roy Hibbert | 86 | 26 | 278 |



Exemple pour $p = 5$: Nous disposons d'un tableau de notes obtenues par $n = 8$ élèves A(riane), B(asile), C(alliope), D(enis), E(ugénie), F(ilippo), G(eorgie), H(ector)

- dans $p = 5$ matières Statistique, Informatique, Marketing, Finance, Comptabilité
 \leadsto les variables x_1, \dots, x_5
- et de leur moyenne pondérée arrondie calculée à partir de ces notes
 \leadsto variable y

| | Statistique | Informatique | Marketing | Finance | Comptabilité | Moyenne_ponderee |
|----------|-------------|--------------|-----------|---------|--------------|------------------|
| A | 13 | 14 | 6 | 8 | 7 | 13.00 |
| B | 16 | 16 | 4 | 8 | 6 | 14.50 |
| C | 6 | 6 | 13 | 15 | 12 | 9.75 |
| D | 7 | 8 | 14 | 16 | 15 | 11.25 |
| E | 16 | 15 | 14 | 14 | 13 | 16.50 |
| F | 17 | 14 | 13 | 15 | 15 | 16.75 |
| G | 6 | 6 | 8 | 7 | 7 | 8.00 |
| H | 7 | 8 | 6 | 6 | 6 | 8.50 |

Si l'on ne sait pas quels coefficients ont été appliqués pour chaque matière, comment les retrouver (approximativement) ?

Formalisation du problème de Régression Linéaire au sens des Moindres Carrés (MC)

Cette méthode est dite

Régression Linéaire au sens des Moindres Carrés (= least squares method)

lorsqu'en plus d'imposer que la fonction f soit affine :

$$f : \mathbf{x}_. = (x_1, \dots, x_p) \in \mathbb{R}^p \mapsto \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad \text{où } \beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}$$

on impose le critère

$$\left(\forall i \in \llbracket 1, n \rrbracket, f(\mathbf{x}_{i.}) \simeq y_i \right) \Leftrightarrow \left(\sum_{i=1}^n (y_i - f(\mathbf{x}_{i.}))^2 \text{ le + petit possible} \right)$$

c-à-d que l'on impose de choisir $(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}$ tel que

$$\sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \right)^2 \text{ soit le plus petit possible}$$

Lorsque

- $p > 1$, on parle de Régression Linéaire **Multiple** au sens des Moindres Carrés
- $p = 1$, on parle de Régression Linéaire **Simple** au sens des Moindres Carrés, et cette méthode revient à imposer, en plus du fait que la fonction f soit affine

$$f : x \in \mathbb{R} \mapsto \beta_0 + \beta_1 x \quad \text{où } \beta_0, \beta_1 \in \mathbb{R}$$

le critère

$$\left(\forall i \in \llbracket 1, n \rrbracket, f(x_i) \simeq y_i \right) \Leftrightarrow \left(\sum_{i=1}^n (y_i - f(x_i))^2 \text{ le + petit possible} \right)$$

c-à-d que l'on impose de choisir $(\beta_0, \beta_1) \in \mathbb{R}^2$ tel que

$$\sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2 \text{ soit le plus petit possible}$$

A partir de maintenant nous dirons juste RLS (resp. RLM) à la place de Régression Linéaire Simple (resp. Multiple) au sens des Moindres Carrés

Réécriture du problème d'optimisation

On cherche donc le $(p+1)$ -uplet $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ qui minimise la somme des **erreurs quadratiques** ϵ_i^2 , où $\epsilon_i := y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$:

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) &= \underset{(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}}{\operatorname{Argmin}} \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \right)^2 \\ &= \underset{(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}}{\operatorname{Argmin}} \sum_{i=1}^n \epsilon_i^2\end{aligned}$$

- Pr $i \in \llbracket 1, n \rrbracket$, $\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip} \rightsquigarrow$ **valeur prédite pour y_i**
- Pr $i \in \llbracket 1, n \rrbracket$, $\hat{\epsilon}_i := y_i - \hat{y}_i \rightsquigarrow$ **résidu dans la prédiction de y_i par \hat{y}_i**

$$\begin{aligned}\bullet \quad & \min_{(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}} \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \right)^2 \\ &= \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}) \right)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 \\ &\rightsquigarrow \text{somme des carrés des résidus (scr)}\end{aligned}$$

Réécriture du problème d'optimisation : cas $p = 1$

On cherche ici le couple $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimise la somme des **erreurs quadratiques** ϵ_i^2 , où $\epsilon_i := y_i - (\beta_0 + \beta_1 x_i)$:

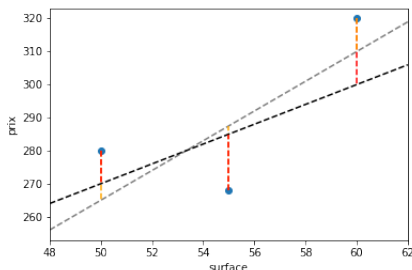
$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1) &= \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{Argmin}} \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2 \\ &= \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{Argmin}} \sum_{i=1}^n \epsilon_i^2\end{aligned}$$

- Pr $i \in \llbracket 1, n \rrbracket$, $\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \rightsquigarrow$ **valeur prédite pour y_i**
- Pr $i \in \llbracket 1, n \rrbracket$, $\hat{\epsilon}_i := y_i - \hat{y}_i \rightsquigarrow$ **résidu dans la prédiction de y_i par \hat{y}_i**
- $$\begin{aligned}\min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2 \\ = \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 \\ \rightsquigarrow \text{somme des carrés des résidus (scr)}\end{aligned}$$

Interprétation graphique de la RL (au sens des MC)

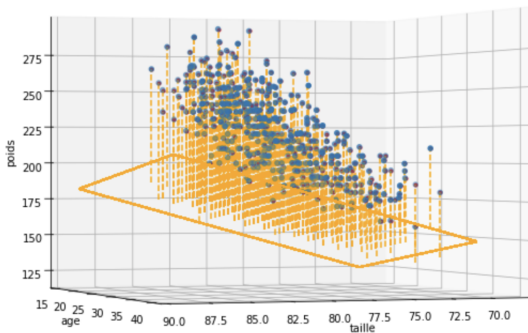
- cas $p = 1$ (RLS) : $\hat{\beta}_0, \hat{\beta}_1$ sont construits pour minimiser les "distances verticales", au sens portées par des droites parallèles à (Oy) , entre les observations (x_i, y_i) et la droite de régression théorique $y = \beta_0 + \beta_1 x$.

Exemple pour $p = 1$: pour les données $(50, 280)$, $(55, 268)$, $(60, 320)$ extraites de nos données appartements = (surfaces, prix), et deux droites quelconques du type $y = b_0 + b_1 x$



- cas $p = 2$: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ sont construits pour minimiser les "distances verticales", au sens portées par des droites parallèles à (Oz) , entre les observations (x_{i1}, x_{i2}, y_i) et le plan affine de régression théorique $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Exemple pour nos données nba avec un plan quelconque :



- cas $p \geq 3$: représentation graphique impossible !

Récapitulatif des objectifs

On veut donc :

- estimer de manière optimale $\beta_0, \beta_1, \dots, \beta_p$ à l'aide des données $(x_1. = (x_{11}, \dots, x_{1p}), y_1), \dots, (x_n. = (x_{n1}, \dots, x_{np}), y_n)$
- mesurer l'importance des variables x_1, \dots, x_p dans l'explication de y
- prédire avec précision $\hat{y}_{n+1}^{\text{prev}}$ la "valeur moyenne" de y pour une nouvelle valeur $(x_{n+1,1}, \dots, x_{n+1,p})$ de x_1, \dots, x_p ,

Lorsque $p = 1$ cela devient

- estimer de manière optimale β_0, β_1 à l'aide des données $(x_1, y_1), \dots, (x_n, y_n)$
- mesurer l'importance de la variable x dans l'explication de y
- prédire avec précision $\hat{y}_{n+1}^{\text{prev}}$, la "valeur moyenne" de y pour une nouvelle valeur x_{n+1} de x

A) Résolution du problème de RL au sens des MC

Conditions de dimension & de rang

Ns ns placerons ds tte la suite sous la double condition (A0) suivante :

- $n > p + 1 \rightsquigarrow$ dans nos données il y a plus d'individus que de variables

- $1_{n,1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, x_{.,1}, \dots, x_{.,p}$ forment une famille libre de $p + 1$ vecteurs de \mathbb{R}^n (c-à-d ces $p + 1$ vecteurs sont linéairement indépendants)

Lorsque $p = 1$, la condition est que les vecteurs $1_{n,1}$ et $x_{.,1}$ soient linéairement indépendants, c-à-d $\exists 1 \leq i \neq k \leq n$ tq $x_i \neq x_k$

Lorsque $p > 1$, la condition équivaut à $x_{..} := \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$ de rang $p + 1$

(on dit alors que $x_{..}$ est de **rang plein** puisque $x_{..} \in \mathcal{M}_{n,p+1}(\mathbb{R})$ et $n > p + 1$) ou encore à ${}^t x_{..} x_{..}$ inversible (puisque ${}^t x_{..} x_{..} \in \mathcal{M}_{p+1}(\mathbb{R})$, $\text{Ker}(x_{..}) = \text{Ker}({}^t x_{..} x_{..})$ et, d'après le Théorème du rang, $p + 1 = \text{rg}(x_{..}) + \dim(\text{Ker}(x_{..})) = \text{rg}({}^t x_{..} x_{..}) + \dim(\text{Ker}({}^t x_{..} x_{..}))$)

Résolution du problème d'optimisation : cas $p = 1$

On veut résoudre le problème d'optimisation de la RLS au sens des MC :

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{Argmin}} C(\beta_0, \beta_1)$$

$$\text{où } C : (\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R} \mapsto \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

\leadsto C peut être vue comme une fonction modélisant le coût d'utilisation de la droite de régression théorique $y = \beta_0 + \beta_1 x$

- C est une fonction deux fois dérivable sur $\mathbb{R} \times \mathbb{R}$
- Recherche des points critiques de C c-à-d des points $(\beta_0, \beta_1) \in \mathbb{R}^2$ tel que $\nabla C(\beta_0, \beta_1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ où $\nabla C(\beta_0, \beta_1)$ est le gradient de C en (β_0, β_1) :

$$\nabla C(\beta_0, \beta_1) = \begin{pmatrix} \frac{\partial C}{\partial \beta_0}(\beta_0, \beta_1) \\ \frac{\partial C}{\partial \beta_1}(\beta_0, \beta_1) \end{pmatrix} = \begin{pmatrix} -2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) \\ -2 \sum_{i=1}^n x_i (y_i - (\beta_0 + \beta_1 x_i)) \end{pmatrix}$$

Cela revient à résoudre le syst. de 2 équations à 2 inconnues (β_0, β_1) :

$$\begin{cases} -2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ -2 \sum_{i=1}^n x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \end{cases}$$

On trouve un unique point critique, que l'on note $(\hat{\beta}_0, \hat{\beta}_1)$:

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \end{cases}$$

Rq : $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\sigma}_x^2 \geq 0$ et est $\hat{\sigma}_x^2 = 0$ ssi $x_1 = x_2 = \dots = x_n$, ce qui est impossible puisqu'on a supposé qu'il existe $1 \leq i \neq k \leq n$ tq $x_i \neq x_k$, donc $\hat{\sigma}_x^2 > 0$

- On détermine la Hessienne de C en $(\hat{\beta}_0, \hat{\beta}_1)$:

$$\text{Hess}C(\hat{\beta}_0, \hat{\beta}_1) = \begin{pmatrix} \frac{\partial^2 C}{\partial \beta_0^2}(\hat{\beta}_0, \hat{\beta}_1) & \frac{\partial^2 C}{\partial \beta_0 \partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) \\ \frac{\partial^2 C}{\partial \beta_0 \partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) & \frac{\partial^2 C}{\partial \beta_1^2}(\hat{\beta}_0, \hat{\beta}_1) \end{pmatrix} = \begin{pmatrix} 2n & 2 \sum_{i=1}^n x_i \\ 2 \sum_{i=1}^n x_i & 2 \sum_{i=1}^n x_i^2 \end{pmatrix}$$

\leadsto elle est définie positive (son déterminant est égal à $4n^2 \hat{\sigma}_x^2$ qui est strictement positif car les x_i ne sont pas tous égaux, et sa trace est égale à $2n(1 + \hat{\sigma}_x^2 + (\bar{x})^2)$ qui est strictement positive) donc C atteint bien un minimum local en $(\hat{\beta}_0, \hat{\beta}_1)$ (en fait, pour tout $(\beta_0, \beta_1) \in \mathbb{R}^2$,

$$\text{Hess}C(\beta_0, \beta_1) = \begin{pmatrix} 2n & 2 \sum_{i=1}^n x_i \\ 2 \sum_{i=1}^n x_i & 2 \sum_{i=1}^n x_i^2 \end{pmatrix} \text{ donc } C \text{ est convexe sur } \mathbb{R}^2, \text{ et } C \text{ atteint bien son minimum } \underline{\text{global}} \text{ en } (\hat{\beta}_0, \hat{\beta}_1)$$

Conclusion pour la RLS : La droite de régression linéaire au sens des

moindres carrés est $y = \hat{\beta}_0 + \hat{\beta}_1 x$ avec
$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\hat{c}_{x,y}}{\hat{\sigma}_x^2} = \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \hat{\rho}_{x,y} \end{cases}$$

où $\hat{c}_{x,y} := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ est la **covariance** de x et y et $\hat{\rho}_{x,y} := \frac{\hat{c}_{x,y}}{\hat{\sigma}_x \hat{\sigma}_y}$ est le **coefficient de corrélation linéaire** de x et y (aussi dit **de Pearson**)

$$\hat{\sigma}_{x,y} \rightsquigarrow \mathbf{x}.\text{cov}(\mathbf{y}, \text{ddof}=0) \text{ ou } \text{np.cov}(\mathbf{x}, \mathbf{y}, \text{ddof}=0)[0,1]$$

$$\hat{\rho}_{x,y} \rightsquigarrow \mathbf{x}.\text{corr}(\mathbf{y}) \text{ ou } \text{np.corrcoef}(\mathbf{x}, \mathbf{y})[0,1]$$

- Comme $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, cette droite passe par le point de coordonnées (\bar{x}, \bar{y}) càd le centre de gravité du nuage de points (aussi appelé point moyen, centre d'inertie, isobarycentre du nuage de points)
- Comme $\hat{\sigma}_x, \hat{\sigma}_y > 0$, le coefficient directeur $\hat{\beta}_1$ de cette droite et $\hat{\rho}_{x,y}$ sont de même signe : la fonction associée est croissante (resp. décroissante) ssi $\hat{\rho}_{x,y}$ est positif (resp. négatif) Cela permet de deviner le signe de $\hat{\rho}_{x,y}$ avec la silhouette du nuage de points $(x_i, y_i)_{i \in \llbracket 1, n \rrbracket}$!

- $\hat{\rho}_{x,y}$ mesure à quel point x et y sont liées linéairement ; $\hat{\rho}_{x,y} \in [-1, 1]$ (conséquence de l'inégalité de Cauchy-Schwarz)

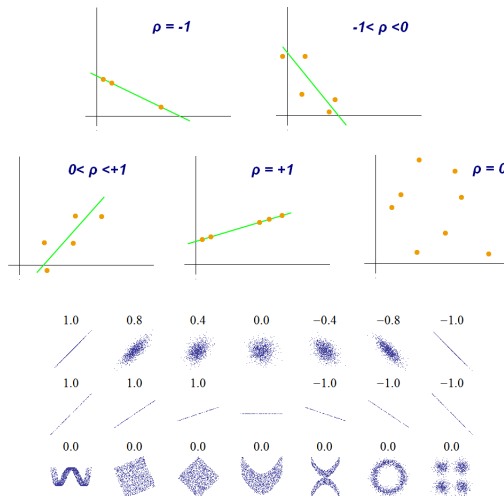
et, plus $|\hat{\rho}_{x,y}|$ est proche de 1, plus la liaison linéaire entre x et y est forte.

$$\text{Rq : } \hat{\sigma}_{x,x} = \hat{\sigma}_x^2 \text{ et } \hat{\rho}_{x,x} = 1$$

$$\text{Pour } \alpha, \alpha', \beta, \beta', \gamma, \gamma' \in \mathbb{R}, \mathbf{u}, \mathbf{v} \in \mathbb{R}^n,$$

$$\hat{\sigma}_{\alpha x + \beta \mathbf{u} + \gamma, \alpha' y + \beta' \mathbf{v} + \gamma'} = \alpha \alpha' \hat{\sigma}_{x,y} + \alpha \beta' \hat{\sigma}_{x,\mathbf{v}} + \beta \alpha' \hat{\sigma}_{\mathbf{u},y} + \beta \beta' \hat{\sigma}_{\mathbf{u},\mathbf{v}}$$

Les graphiques ci-dessous illustrent le lien existant entre la pertinence de l'ajustement d'un nuage de points par une droite, caractérisée par la corrélation linéaire entre x et y , et son coefficient $\hat{\rho}_{x,y}$:



Résolution du problème d'optimisation : cas $p > 1$

En notant : $y_{\cdot} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $x_{\cdot\cdot} := \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$, $\beta_{\cdot} := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, nous avons

$$C(\beta_{\cdot}) = C(\beta_0, \beta_1, \dots, \beta_p) := \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \right)^2 = \|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$$

On cherche donc : $\hat{\beta}_{\cdot} = \underset{\beta_{\cdot} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} C(\beta_0, \beta_1, \dots, \beta_p) = \underset{\beta_{\cdot} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$


ce qui équivaut à chercher $\underset{z_{\cdot} \in \operatorname{Im}(x_{\cdot\cdot})}{\operatorname{argmin}} \|y_{\cdot} - z_{\cdot}\|_{\mathbb{R}^n}^2$

• Si $\hat{\beta}_{\cdot} = \underset{\beta_{\cdot} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$, on a : $\forall \beta_{\cdot} \in \mathbb{R}^{p+1}$, $\|y_{\cdot} - x_{\cdot\cdot} \hat{\beta}_{\cdot}\|_{\mathbb{R}^n}^2 \leq \|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$ ce qui est équivalent (puisque $\operatorname{Im}(x_{\cdot\cdot})$)

est par déf. l'ensemble des $x_{\cdot\cdot} \beta_{\cdot}$ avec β_{\cdot} variant dans \mathbb{R}^{p+1}), en notant $\hat{y}_{\cdot} = x_{\cdot\cdot} \hat{\beta}_{\cdot}$ (qui par déf. $\in \operatorname{Im}(x_{\cdot\cdot})$) à $\forall z_{\cdot} \in \operatorname{Im}(x_{\cdot\cdot})$, $\|y_{\cdot} - \hat{y}_{\cdot}\|_{\mathbb{R}^n}^2 \leq \|y_{\cdot} - z_{\cdot}\|_{\mathbb{R}^n}^2$, ce qui équivaut aussi à $\hat{y}_{\cdot} = \underset{z_{\cdot} \in \operatorname{Im}(x_{\cdot\cdot})}{\operatorname{argmin}} \|y_{\cdot} - z_{\cdot}\|_{\mathbb{R}^n}^2$

• Inversement, si $z_{\cdot}^* \in \operatorname{Im}(x_{\cdot\cdot})$ est tel que $z_{\cdot}^* = \underset{z_{\cdot} \in \operatorname{Im}(x_{\cdot\cdot})}{\operatorname{argmin}} \|y_{\cdot} - z_{\cdot}\|_{\mathbb{R}^n}^2$, on a $\forall z_{\cdot} \in \operatorname{Im}(x_{\cdot\cdot})$, $\|y_{\cdot} - z_{\cdot}^*\|_{\mathbb{R}^n}^2 \leq \|y_{\cdot} - z_{\cdot}\|_{\mathbb{R}^n}^2$, ce qui

équivaut à $\forall \beta_{\cdot} \in \mathbb{R}^{p+1}$, $\|y_{\cdot} - z_{\cdot}^*\|_{\mathbb{R}^n}^2 \leq \|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$. Comme $z_{\cdot}^* \in \operatorname{Im}(x_{\cdot\cdot})$, $\exists \beta_{\cdot}^* \in \mathbb{R}^{p+1}$ tq $z_{\cdot}^* = x_{\cdot\cdot} \beta_{\cdot}^*$, et l'on a

$\forall \beta_{\cdot} \in \mathbb{R}^{p+1}$, $\|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}^*\|_{\mathbb{R}^n}^2 \leq \|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$ donc $\beta_{\cdot}^* = \underset{\beta_{\cdot} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|y_{\cdot} - x_{\cdot\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$ 

Propriété fondamentale de la RLM (A CONNAITRE PAR COEUR)

Sous la condition double (A0) : $n > p + 1$ et ${}^t x_{..} x_{..}$ inversible
le problème de RLM $\operatorname{argmin}_{\beta_{.} \in \mathbb{R}^{p+1}} \|y_{.} - x_{..} \beta_{.}\|_{\mathbb{R}^n}^2$ admet une unique solution :

$$\hat{\beta}_{.} = ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y_{.}$$

→ $\forall j \in \{0, 1, \dots, p\}$, l'estimateur des moindres carrés de β_j est $\hat{\beta}_j$

→ $\hat{y}_{.} = x_{..} \hat{\beta}_{.} = \hat{\beta}_0 1_{n,1} + \hat{\beta}_1 x_{.,1} + \dots + \hat{\beta}_p x_{.,p} = x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y_{.} = \Pi_{\operatorname{Im}(x_{..})}(y_{.})$

c-à-d que le vecteur des valeurs prédites, $\hat{y}_{.}$, est le projeté orthogonal du vecteur $y_{.}$ des valeurs de la variable à expliquer sur l'espace engendré par les colonnes de $x_{..}$, c-à-d par $1_{n,1}$ (l'intercept) et $x_{.,1}, \dots, x_{.,p}$ les vecteurs contenant les valeurs des variables explicatives

→ l'hyperplan affine de régression linéaire au sens des moindres carrés est $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$

→ $\hat{\epsilon}_{.} = y_{.} - \hat{y}_{.} = (I_n - x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..}) y_{.} = \Pi_{(\operatorname{Im}(x_{..}))^\perp}(y_{.})$ où $\Pi_{(\operatorname{Im}(x_{..}))^\perp}$ est la matrice de projection orthogonale sur $(\operatorname{Im}(x_{..}))^\perp$

Remarque 1 : Cette matrice $x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..}$ de projection orthogonale sur $\text{Im}(x_{..})$ est dite **hat matrix** (ou encore **projection matrix** ou **influence matrix**) car elle met des "hat" sur y . puisque $\hat{y}_{.} = x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y_{.}$
 \leadsto elle envoie le vecteur "valeurs réponses" (= **response values**) sur le vecteur "valeurs prédites" (**fitted values**).

En tant que matrice de projection orthogonale, elle est

- symétrique $\leadsto {}^t (x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..}) = x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..}$
- idempotente $\leadsto (x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..})^2 = (x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..})$

Remarque 2 : on a ${}^t_{X..X..} = \begin{pmatrix} 1 & \dots & 1 \\ x_{11} & \dots & x_{n1} \\ \vdots & \dots & \vdots \\ x_{1p} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$

donc ${}^t_{X..X..} = \begin{pmatrix} n & n\overline{x_{.1}} & n\overline{x_{.2}} & \dots & n\overline{x_{.p}} \\ n\overline{x_{.1}} & \|x_{.1}\|_{\mathbb{R}^n}^2 & \langle x_{.1}, x_{.2} \rangle_{\mathbb{R}^n} & \dots & \langle x_{.1}, x_{.p} \rangle_{\mathbb{R}^n} \\ n\overline{x_{.2}} & \langle x_{.2}, x_{.1} \rangle_{\mathbb{R}^n} & \|x_{.2}\|_{\mathbb{R}^n}^2 & \dots & \langle x_{.2}, x_{.p} \rangle_{\mathbb{R}^n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n\overline{x_{.p}} & \langle x_{.p}, x_{.1} \rangle_{\mathbb{R}^n} & \langle x_{.p}, x_{.2} \rangle_{\mathbb{R}^n} & \dots & \|x_{.p}\|_{\mathbb{R}^n}^2 \end{pmatrix}$

Cela donne lorsque $p = 1$: ${}^t_{X..X..} = \begin{pmatrix} n & n\overline{x_{.}} \\ n\overline{x_{.}} & \|x_{.}\|_{\mathbb{R}^n}^2 \end{pmatrix}$ d'où

$\det({}^t_{X..X..}) = n\|x_{.}\|_{\mathbb{R}^n}^2 - n^2(\overline{x_{.}})^2 = n^2\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2\right) = n^2\hat{\sigma}_{x_{.}}^2$ (qui est > 0 si les

x_i ne sont pas tous égaux) $({}^t_{X..X..})^{-1} = \frac{1}{n^2\hat{\sigma}_{x_{.}}^2} \begin{pmatrix} \|x_{.}\|_{\mathbb{R}^n}^2 & -n\overline{x_{.}} \\ -n\overline{x_{.}} & n \end{pmatrix}$ ce qui permet, avec

$\hat{\beta}_{.} = ({}^t_{X..X..})^{-1} {}^t_{X..y_{.}}$, de retrouver $\begin{cases} \hat{\beta}_0 = \overline{y_{.}} - \hat{\beta}_1\overline{x_{.}} \\ \hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x_{.}})(y_i - \overline{y_{.}})}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x_{.}})^2} \end{cases}$

Preuve n° 1 : Projection orthogonale (1/4)

$\operatorname{argmin}_{z \in \operatorname{Im}(x_{..})} \|y_{.} - z_{.}\|_{\mathbb{R}^n}^2$ est atteint en $z_{.} = \Pi_{\operatorname{Im}(x_{..})}(y_{.}) = \Pi_{\operatorname{Im}(x_{..})} y_{.}$

où $\Pi_{\operatorname{Im}(x_{..})} \in \mathcal{M}_n(\mathbb{R})$ est la matrice de projection orthogonale sur $\operatorname{Im}(x_{..})$
(et également, par abus de notation, l'application linéaire qui lui est canoniquement associée)

↪ donc $\hat{y}_{.} = x_{..} \hat{\beta} = \Pi_{\operatorname{Im}(x_{..})}(y_{.})$

(Rq : on a bien $\hat{y}_{.} \in \mathbb{R}^n$, puisque, d'une part, $x_{..} \in \mathcal{M}_{n,p+1}(\mathbb{R})$ et $\hat{\beta} \in \mathbb{R}^{p+1}$, et d'autre part $\Pi_{\operatorname{Im}(x_{..})} \in \mathcal{M}_n(\mathbb{R})$ et $y_{.} \in \mathbb{R}^n$)

En effet, d'après le Théorème de Pythagore,

$$\begin{aligned} \forall z_{.} \in \operatorname{Im}(x_{..}), \quad \|y_{.} - z_{.}\|_{\mathbb{R}^n}^2 &= \|y_{.} - \Pi_{\operatorname{Im}(x_{..})}(y_{.}) + \Pi_{\operatorname{Im}(x_{..})}(y_{.}) - z_{.}\|_{\mathbb{R}^n}^2 \\ &= \|y_{.} - \Pi_{\operatorname{Im}(x_{..})}(y_{.})\|_{\mathbb{R}^n}^2 + \|\Pi_{\operatorname{Im}(x_{..})}(y_{.}) - z_{.}\|_{\mathbb{R}^n}^2 \end{aligned}$$

car • $y_{.} - \Pi_{\operatorname{Im}(x_{..})}(y_{.}) = \Pi_{(\operatorname{Im}(x_{..}))^\perp}(y_{.}) \in (\operatorname{Im}(x_{..}))^\perp$

• $\Pi_{\operatorname{Im}(x_{..})}(y_{.}) - z_{.} \in \operatorname{Im}(x_{..})$ puisque $\operatorname{Im}(x_{..})$ est un e.v., $\Pi_{\operatorname{Im}(x_{..})}(y_{.}) \in \operatorname{Im}(x_{..})$ et $z_{.} \in \operatorname{Im}(x_{..})$

Donc le minimum est atteint en un unique point :

lorsque $\|\Pi_{\operatorname{Im}(x_{..})}(y_{.}) - z_{.}\|_{\mathbb{R}^n}^2 = 0$, c-à-d lorsque $z_{.} = \Pi_{\operatorname{Im}(x_{..})}(y_{.})$

Preuve n° 1 : Projection orthogonale (2/4)

Notons $z_{\cdot}^* = \Pi_{\text{Im}(x_{\cdot})}(y_{\cdot})$. On a alors $z_{\cdot}^* \in \text{Im}(x_{\cdot})$, il existe donc $\beta_{\cdot}^* \in \mathbb{R}^{p+1}$ tel que $z_{\cdot}^* = x_{\cdot} \beta_{\cdot}^*$, et, comme x_{\cdot} est de rang plein, ce β_{\cdot}^* est unique, (en effet, s'il existait $\tilde{\beta}_{\cdot} \in \mathbb{R}^{p+1}$ tel que $x_{\cdot} \beta_{\cdot}^* = x_{\cdot} \tilde{\beta}_{\cdot}$, alors on aurait $x_{\cdot} (\beta_{\cdot}^* - \tilde{\beta}_{\cdot}) = 0_{n,1}$; or comme x_{\cdot} est supposé de rang plein (c'est-à-dire de rang $p+1$ puisque $x_{\cdot} \in \mathcal{M}_{n,p+1}(\mathbb{R})$ et $n > p+1$), on a d'après le Théorème du rang que $\dim(\text{Ker}(x_{\cdot})) = 0$, d'où $\text{Ker}(x_{\cdot}) = (0_{p+1,1})$, et donc $\beta_{\cdot}^* = \tilde{\beta}_{\cdot}$) et l'on a $\beta_{\cdot}^* = \underset{\beta_{\cdot} \in \mathbb{R}^{p+1}}{\text{argmin}} \|y_{\cdot} - x_{\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$

Avec les notations $\hat{y}_{\cdot} = z_{\cdot}^* = \Pi_{\text{Im}(x_{\cdot})}(y_{\cdot})$ et $\hat{\beta}_{\cdot} = \beta_{\cdot}^*$ on a que $\hat{y}_{\cdot} = x_{\cdot} \hat{\beta}_{\cdot}$ et $\hat{\beta}_{\cdot}$ est l'unique solution de $\underset{\beta_{\cdot} \in \mathbb{R}^{p+1}}{\text{argmin}} \|y_{\cdot} - x_{\cdot} \beta_{\cdot}\|_{\mathbb{R}^n}^2$

Il reste à déterminer l'expression explicite :

↪ de la solution $\hat{\beta}_{\cdot}$ du pb et de \hat{y}_{\cdot}

↪ de la matrice $\Pi_{\text{Im}(x_{\cdot})}$ de projection orthogonale sur l'espace engendré par les colonnes de la matrice x_{\cdot}

Preuve n° 1 : Projection orthogonale (3/4)

Méthode 1 :

$$\begin{aligned} \forall v \in \text{Im}(x_{..}) , \quad \langle v, (I_n - \Pi_{\text{Im}(x_{..})}) y. \rangle &= 0 \\ \Leftrightarrow \forall u \in \mathbb{R}^{p+1} , \quad \langle x_{..} u, (I_n - \Pi_{\text{Im}(x_{..})}) y. \rangle &= 0 \\ \Leftrightarrow \forall u \in \mathbb{R}^{p+1} , \quad {}^t u. {}^t x_{..} (I_n - \Pi_{\text{Im}(x_{..})}) y. &= 0 \\ \Leftrightarrow {}^t x_{..} (I_n - \Pi_{\text{Im}(x_{..})}) y. &= 0_{p+1,1} \\ \Leftrightarrow {}^t x_{..} y. &= {}^t x_{..} \Pi_{\text{Im}(x_{..})} y. \\ \Leftrightarrow {}^t x_{..} y. &= {}^t x_{..} x_{..} \hat{\beta}. \\ \Leftrightarrow ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y. &= \hat{\beta}. \quad \text{car } {}^t x_{..} x_{..} \text{ inversible} \end{aligned}$$

Méthode 2 :

$$\begin{aligned} (I_n - \Pi_{\text{Im}(x_{..})}) y. \in (\text{Im}(x_{..}))^\perp &\Leftrightarrow \forall j \in \llbracket 0, p \rrbracket, x_{.j} \perp (I_n - \Pi_{\text{Im}(x_{..})}) y. \\ \Leftrightarrow \forall j \in \llbracket 0, p \rrbracket, x_{.j} &\perp (y. - x_{..} \hat{\beta}.) \\ \Leftrightarrow {}^t x_{..} y. &= {}^t x_{..} x_{..} \hat{\beta}. \Leftrightarrow \hat{\beta}. = ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y. \quad \text{car } {}^t x_{..} x_{..} \text{ inversible} \end{aligned}$$

Preuve n° 1 : Projection orthogonale (4/4)

Conclusion \leadsto Détermination de l'expression explicite de $\Pi_{\text{Im}(x_{..})}$:

Comme $\hat{\beta}_{.} = ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y_{.}$ et $\hat{y} = x_{..} \hat{\beta}_{.} = \Pi_{\text{Im}(x_{..})} y_{.}$, on a

$$x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y_{.} = \Pi_{\text{Im}(x_{..})} y_{.}$$

Et comme ceci est vrai pour tout $y_{.} \in \mathbb{R}^n$, on a finalement

$$\Pi_{\text{Im}(x_{..})} = x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..}$$

Rq : $x_{..}$ est de taille $n \times (p+1)$, donc ${}^t x_{..}$ est de tailles $(p+1) \times n$, et ${}^t x_{..} x_{..}$ est de taille $(p+1) \times (p+1)$, de même que $({}^t x_{..} x_{..})^{-1}$ d'où finalement $({}^t x_{..} x_{..})^{-1} {}^t x_{..}$ est de taille $(p+1) \times n$ et $x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..}$ est de taille $n \times n$; enfin comme $y_{.}$ est de taille $n \times 1$, $\hat{\beta}_{.}$ est de taille $(p+1) \times 1$

Preuve n° 2 : Optimisation

On cherche $\hat{\beta}_. = \operatorname{argmin}_{\beta_. \in \mathbb{R}^{p+1}} C(\beta_0, \beta_1, \dots, \beta_p)$ où $C(\beta_0, \beta_1, \dots, \beta_p) = \|y - x_.. \beta_.\|_{\mathbb{R}^n}^2$

- Condition d'ordre 1 c-à-d recherche des points critiques de C
c-à-d des points $\beta_. = (\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}$ annulant le gradient de C

$$\text{c-à-d tel que } \nabla C(\beta_.) = \nabla C(\beta_0, \beta_1, \dots, \beta_p) = \begin{pmatrix} \frac{\partial C}{\partial \beta_0}(\beta_0, \beta_1, \dots, \beta_p) \\ \frac{\partial C}{\partial \beta_1}(\beta_0, \beta_1, \dots, \beta_p) \\ \vdots \\ \frac{\partial C}{\partial \beta_p}(\beta_0, \beta_1, \dots, \beta_p) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0_{p+1,1}$$

Méthode 1 \leadsto différentiabilité : comme C différentiable sur \mathbb{R}^{p+1} et

- $\forall h_. \in \mathbb{R}^{p+1}, \|y - x_..(\beta_. + h_.)\|_{\mathbb{R}^n}^2 = \|y - x_.. \beta_.\|_{\mathbb{R}^n}^2 - 2\langle y - x_.. \beta_., x_.. h_. \rangle_{\mathbb{R}^n} + \|x_.. h_.\|_{\mathbb{R}^n}^2$
- $h_. \in \mathbb{R}^{p+1} \mapsto -\langle 2(y - x_.. \beta_.), x_.. h_. \rangle_{\mathbb{R}^n} = -\langle 2 {}^t x_.. (y - x_.. \beta_.), h_. \rangle_{\mathbb{R}^{p+1}}$ linéaire
- $\|x_.. h_.\|_{\mathbb{R}^n}^2 = o(\|h_.\|_{\mathbb{R}^{p+1}})$

on a par identification que $\nabla C(\beta_.) = -2 {}^t x_.. (y - x_.. \beta_.)$, et donc les points critiques $\beta_.$ de C sont caractérisés par ${}^t x_.. y = {}^t x_.. x_.. \beta_.$, c-à-d, comme ${}^t x_.. x_..$ inversible, par $\beta_. = ({}^t x_.. x_..)^{-1} {}^t x_.. y$.

Preuve n° 2 : Optimisation

Méthode 2 \leadsto calcul composante par composante de $\nabla C(\beta.)$

$$\begin{aligned}\|y. - x_{..}\beta.\|_{\mathbb{R}^n}^2 &= {}^t(y. - x_{..}\beta.) (y. - x_{..}\beta.) = ({}^ty. - {}^t\beta. {}^tx_{..}) (y. - x_{..}\beta.) \\ &= {}^ty. y. - {}^ty. x_{..} \beta. - {}^t\beta. {}^tx_{..} y. + {}^t\beta. {}^tx_{..} x_{..}\beta. \\ &= {}^ty. y. - 2 {}^t\beta. {}^tx_{..} y. + {}^t\beta. {}^tx_{..} x_{..}\beta.\end{aligned}$$

car ${}^t(x_{..}\beta.) = {}^t\beta. {}^tx_{..}$ et, comme ${}^ty. x_{..} \beta.$ est un réel, il est égal à sa transposée c-à-d ${}^ty. x_{..} \beta. = {}^t\beta. {}^tx_{..} y.$

Notons (e_0, e_1, \dots, e_p) la base canonique de \mathbb{R}^{p+1} (c-à-d $\forall j \in \{0, 1, \dots, p\}$, e_j est le vecteur colonne à $p+1$ composantes avec p composantes nulles sauf la $j+1$ -ème qui vaut 1 ; rq : on a donc pr tt $u = (u_0, u_1, \dots, u_p) \in \mathbb{R}^{p+1}$, ${}^te_j u = u_j$ est la $(j+1)$ -ème composante de u)

Soit $j \in \{0, 1, \dots, p\}$, déterminons $\frac{\partial C}{\partial \beta_j}(\beta.)$: par linéarité de la dérivation, on a

$$\begin{aligned}\frac{\partial C}{\partial \beta_j}(\beta.) &= \frac{\partial(\beta. \mapsto {}^ty. y.)}{\partial \beta_j}(\beta.) + \frac{\partial(\beta. \mapsto -2 {}^t\beta. {}^tx_{..} y.)}{\partial \beta_j}(\beta.) + \frac{\partial(\beta. \mapsto {}^t\beta. {}^tx_{..} x_{..}\beta.)}{\partial \beta_j}(\beta.) \\ &= 0 - 2 {}^te_j {}^tx_{..} y. + {}^te_j {}^tx_{..} x_{..}\beta. + {}^t\beta. {}^tx_{..} x_{..}e_j \\ &= -2 {}^te_j {}^tx_{..} y. + 2 {}^te_j {}^tx_{..} x_{..}\beta.\end{aligned}$$

car ${}^te_j {}^tx_{..} x_{..}\beta.$ est un réel, donc il est égal à sa transposée, d'où ${}^te_j {}^tx_{..} x_{..}\beta. = {}^t\beta. {}^tx_{..} x_{..}e_j$

$$\begin{aligned}\text{Donc } \nabla C(\beta.) = 0 &\Leftrightarrow \forall j \in \{0, 1, \dots, p\}, \frac{\partial C}{\partial \beta_j}(\beta.) = 0 \Leftrightarrow -2 {}^te_j {}^tx_{..} y. + \\ 2 {}^te_j {}^tx_{..} x_{..}\beta. &= 0 \Leftrightarrow {}^te_j {}^tx_{..} x_{..}\beta. = {}^te_j {}^tx_{..} y. \Leftrightarrow \square {}^tx_{..} x_{..}\beta. \stackrel{=}{=} {}^tx_{..} y.\end{aligned}$$

- Condition d'ordre deux : montrons que la fct C est convexe c-à-d que sa Hessienne est définie positive, ce qui impliquera que le point critique (unique) $\hat{\beta}_. = \left({}^t x_{..} x_{..} \right)^{-1} {}^t x_{..} y_.$ est le minimum (unique) de C

Notons $HC(\beta_.)$ la matrice Hessienne de C au point $\beta_.$:

- $\forall \beta_. \in \mathbb{R}^{p+1}$, $HC(\beta_.) = 2 {}^t x_{..} x_{..}$ (et donc HC est constante) : $\forall j, k \in \{0, 1, \dots, p\}$

$$\begin{aligned} \frac{\partial^2 C}{\partial \beta_j \partial \beta_k}(\beta_.) &= \frac{\partial \left(\frac{\partial C}{\partial \beta_j} \right)}{\partial \beta_k}(\beta_.) = \frac{\partial \left(\beta_. \mapsto -2 {}^t e_j {}^t x_{..} y_. + 2 {}^t e_j {}^t x_{..} x_{..} \beta_. \right)}{\partial \beta_k}(\beta_.) \\ &= -2 \frac{\partial \left(\beta_. \mapsto {}^t e_j {}^t x_{..} y_. \right)}{\partial \beta_k} + 2 \frac{\partial \left(\beta_. \mapsto 2 {}^t e_j {}^t x_{..} x_{..} \beta_. \right)}{\partial \beta_k} \\ &= 0 + 2 {}^t e_j {}^t x_{..} x_{..} e_{.k} = 2 {}^t e_j {}^t x_{..} x_{..} e_{.k} \end{aligned}$$

- HC est définie positive : pour tout $z_. = (z_0, z_1, \dots, z_p) \in \mathbb{R}^{p+1}$ non nul, ${}^t z_. HC(\beta_.) z_. = {}^t z_. \left(2 {}^t x_{..} x_{..} \right) z_. = 2 \|x_{..} z_.\|_{\mathbb{R}^{p+1}}^2 > 0$. En effet,

* $\|x_{..} z_.\|_{\mathbb{R}^{p+1}}^2 \geq 0$ en tant que norme au carré

* $\|x_{..} z_.\|_{\mathbb{R}^{p+1}}^2 = 0$ si et seulement si $x_{..} z_. = 0_{n,1}$ ce qui est impossible puisque,

comme $x_{..}$ est supposé de rang $p+1$, on a d'après le Théorème du rang que $\dim(\text{Ker}(x_{..})) = 0$, d'où $\text{Ker}(x_{..}) = (0_{p+1,1})$, or

$z_. \neq 0_{p+1,1}$ par hypothèse donc $\|x_{..} z_.\|_{\mathbb{R}^{p+1}}^2 \neq 0$

Remarque : comme $\partial_{\beta_0} C(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = 0$, on a

$$\Leftrightarrow \sum_{i=1}^n y_i = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})$$

$$\Leftrightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x_{.1}} + \dots + \hat{\beta}_p \overline{x_{.p}}$$

$$\Leftrightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x_{.1}} + \dots + \hat{\beta}_p \overline{x_{.p}}$$

autrement dit le centre de gravité $(\overline{x_{.1}}, \dots, \overline{x_{.p}}, \bar{y})$ appartient à l'hyperplan affine solution au sens des moindres carrés.

Application numérique & graphique : cas $p = 1$

* avec les formules explicites :

```
hatbeta1 = np.cov(x, y, ddof=0)[0, 1] / np.var(x)
hatbeta0 = np.mean(y) - hatbeta1 * np.mean(x)
```

* avec `lstsq` (*LeaST Squares*) de `scipy.linalg` :

```
n = len(x)
xaugm = np.ones((n,2))
xaugm[:,1] = x
hatbeta = linalg.lstsq(xaugm, y)[0]
```

* avec `LinearRegression` de `sklearn.linearmodel` :

```
regmod = linearmodel.LinearRegression().fit(xaugm, y)
hatbeta0 = regmod.intercept
hatbeta1 = regmod.coef[0]
```

* avec OLS (*Ordinary Least Squares*) de `statsmodels` :

```
xaugm = statsmodels.api.add_constant(x)
modelelin = statsmodels.api.OLS(y, xaugm)
reg = modelelin.fit()
hatbeta = reg.params
```

On trouve pour l'exemple des appartements :

$$\hat{\beta}_0 \simeq 33.64381565 \simeq 33.644$$

$$\hat{\beta}_1 \simeq 3.84782015 \simeq 3.848$$

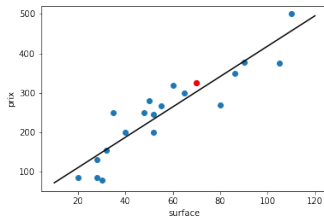
$$\text{c-à-d : } \text{prix} \simeq 33.644 + 3.848 \times \text{surface}$$

(en noir : la droite $y = \hat{\beta}_0 + \hat{\beta}_1 x$;

en bleu : les appartements ;

en rouge : un point de repère,

le 14^{ème} appartement (70, 325))



Application numérique & graphique : cas $p = 2$

```
x_tab = nba_tab[["taille", "age"]]  
n = nba.index.size  
x_tab["intercept"] = np.ones((n,1))  
y_vect = nba_tab["poids"]
```

* avec les formules matricielles explicites :

```
hat_beta = np.dot(linalg.inv(np.dot(np.transpose(x_tab),  
                                     x_tab)), np.dot(np.transpose(x_tab), y_vect))
```

* avec lstsq de scipy.linalg :

```
hat_beta = linalg.lstsq(x_tab, y_vect)[0]
```

* avec sklearn.linear_model.LinearRegression :

```
reg_mod = LinearRegression().fit(x_tab, y_vect)  
hat_beta_0 = reg_mod.intercept_  
hat_beta_1, hat_beta_2 = reg_mod.coef_[0:2]
```

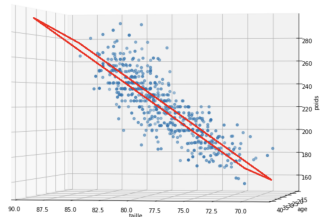
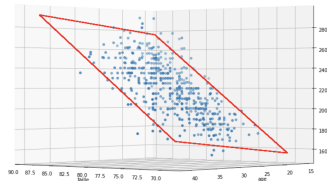
* avec OLS (*Ordinary Least Squares* de statsmodels) :

```
modelelin = statsmodels.api.OLS(y_vect, x_tab)  
reg = modelelin.fit()  
hat_beta = reg.params
```

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

où $\hat{\beta}_0 \simeq -296.599$, $\hat{\beta}_1 \simeq 6.327$, $\hat{\beta}_2 \simeq 0.651$

c-à-d : $\text{poids} \simeq -296.599 + 6.327 * \text{taille} + 0.651 * \text{âge}$



Application numérique & graphique : cas $p = 5$

| | Statistique | Informatique | Marketing | Finance | Comptabilité | Moyenne_ponderee |
|---|-------------|--------------|-----------|---------|--------------|------------------|
| A | 13 | 14 | 6 | 8 | 7 | 13.00 |
| B | 16 | 16 | 4 | 8 | 6 | 14.50 |
| C | 6 | 6 | 13 | 15 | 12 | 9.75 |
| D | 7 | 8 | 14 | 16 | 15 | 11.25 |
| E | 16 | 15 | 14 | 14 | 13 | 16.50 |
| F | 17 | 14 | 13 | 15 | 15 | 16.75 |
| G | 6 | 6 | 8 | 7 | 7 | 8.00 |
| H | 7 | 8 | 6 | 6 | 6 | 8.50 |

```
matieres = tab_notes.columns[0:5]
x_tab = tab_notes[matieres]
x_tab['intercept'] = np.ones((n,1))
y_vect = tab_notes['Moyenne_ponderee']
hat_beta = np.dot(linalg.inv(np.dot(np.transpose(x_tab), x_tab)) , np.dot(np.transpose(x_tab), y_vect) )
ce qui donne : [0.3938768 0.30034548 0.1089102 0.10380356 0.08649371 1.5899103 ]
```

```
ou avec sklearn.linear_model.LinearRegression
matieres = tab_notes.columns[0:5]
x_tab = tab_notes[matieres]
reg = LinearRegression().fit(x_tab, y_vect)
print(reg.coef_)    ce qui donne : [0.3938768 0.30034548 0.1089102 0.10380356 0.08649371 0. ]
print(reg.intercept_) ce qui donne : 1.5899102986314961
```

Rq : les vrais coefficients étaient $[0.4, 0.3, 0.1, 0.1, 0.1]$ et 1.5

Deux questions supplémentaires :

- Comment utiliser ces résultats pour faire des prévisions ?
- Comment évaluer l'adéquation du modèle aux données ?

Prévision de l'étiquette d'une nouvelle donnée

Supposons que l'on dispose d'une nouvelle donnée, non étiquetée, $(x_{n+1,1}, \dots, x_{n+1,p})$, mais qu'a priori l'on peut supposer son comportement "proche" de celui des données déjà connues.

Autrement dit, l'étiquette y_{n+1} de x_{n+1} doit vérifier $y_{n+1} \simeq \beta_0 + \sum_{j=1}^p \beta_j x_{n+1,j}$

On utilise alors comme prévision de y_{n+1} la valeur

$$\hat{y}_{n+1}^{\text{prev}} := \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{n+1,j} = x_{n+1,.} \hat{\beta}. \quad \text{avec } x_{n+1,.} := (1, x_{n+1,1}, \dots, x_{n+1,p}).$$

Ex pr $p = 2$ (basketteurs) : comme $\hat{\beta}_0 \simeq -296.599$, $\hat{\beta}_1 \simeq 6.327$, $\hat{\beta}_2 \simeq 0.651$, pour un nouveau basketteur, de taille $x_{(n+1),1}^{\text{prev}} = 88$ et d'âge $x_{(n+1),2}^{\text{prev}} = 20$, on peut supposer qu'il pèsera

$$\hat{y}_{n+1}^{\text{prev}} := \hat{\beta}_0 + \hat{\beta}_1 x_{(n+1),1}^{\text{prev}} + \hat{\beta}_2 x_{(n+1),2}^{\text{prev}} \simeq -296.599 + 6.327 * 88 + 0.651 * 20 \simeq 273$$

avec OLS

```
taille_new, age_new = 88, 20
```

```
x_new = np.array([1, taille_new, age_new])
```

```
y_prev = reg.predict(x_new)
```

Cas de la RLS : supposons que l'on dispose d'une nouvelle donnée, non étiquetée, x_{n+1} , mais qu'a priori l'on peut supposer son comportement "proche" de celui des données déjà connues.

Autrement dit, l'étiquette y_{n+1} de x_{n+1} doit vérifier $y_{n+1} \simeq \beta_0 + \beta_1 x_{n+1}$

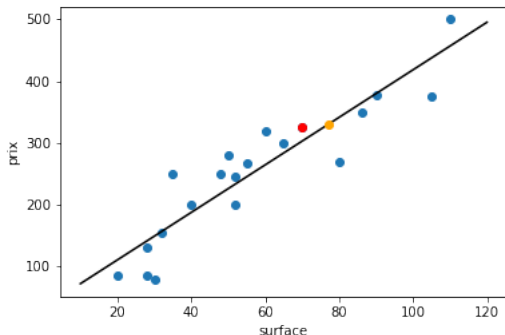
On utilise alors comme prévision de y_{n+1} la valeur

$$\hat{y}_{n+1}^{\text{prev}} := \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$

Ex pr $p = 1$ (appartements) : en postulant que le nouvel appartement, de surface $x_{n+1} = 77$, que notre agent prend en charge suit la même tendance que les appartements du même quartier qu'il a déjà vendus, autrement dit, que son étiquette y_{n+1} vérifie

$y_{n+1} \simeq \beta_0 + \beta_1 x_{n+1}$, notre agent peut fixer le prix de vente de ce nouvel appartement à

$$\hat{y}_{n+1}^{\text{prev}} := \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \simeq 33.644 + 3.848 * 77 \simeq 330 \quad (\text{point orange})$$



```
surface21 = 77
```

```
yprev21 = hatbeta0 + hatbeta1 * surface21
```

ou bien, en utilisant OLS

```
surface21augm = np.array([1, surface21])
```

```
yprev21 = reg.predict(surface21augm)
```

ou encore, en utilisant les notations matricielles

```
yprev21 = np.dot(surface21augm, hatbetam)
```


Evaluation de l'adéquation du modèle ANOVA (Analyse de la Variabilité) & R^2

- **Somme des Carrés Totale** $sct := \|y. - \bar{y}.1_{n,1}\|_{\mathbb{R}^n}^2 = \sum_{i=1}^n (y_i - \bar{y}.)^2 = n\hat{\sigma}_y^2$
- **Somme des Carrés Expliquée** $sce := \|\hat{y}. - \bar{y}.1_{n,1}\|_{\mathbb{R}^n}^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}.)^2$
- **Somme des Carrés Résiduelle** $scr := \|y. - \hat{y}.\|_{\mathbb{R}^n}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2$

Rq : d'après le Théorème de Pythagore, $sct = sce + scr$ (exercice !)

- **Coefficient de détermination** $r^2 := \frac{sce}{sct} = 1 - \frac{scr}{sct}$
 $\leadsto r^2 \in [0, 1]$ correspond à la proportion de variabilité des y_i expliquée
par le modèle linéaire $y = \beta_0 + \sum_{j=1}^p \beta_j x_j$

Plus r^2 est proche de 1, plus le modèle est adapté aux données

Remarque : lorsque $p = 1$, on a $r^2 = \hat{\rho}_{x,y}^2$

Exemple pour $p = 1$ (appartements) :

```
hatprix = hatbeta0 + hatbeta1 * np.array(surface)
hatepsilon = np.array(prix) - hatprix
scr = np.sum(hatepsilon**2)
sct = n * np.var(prix)
sce = np.sum( ( hatprix - np.mean(prix) )**2)
r2 = sce/sct
```

ou bien, avec la fonction OLS :

```
hatprix = reg.fittedvalues
hatepsilon = reg.resid
scr = reg.ssr
sce = reg.ess
sct = reg.centered_tss
r2 = reg.rsquared
```

ou encore, avec les notations matricielles

```
hatprixm = np.dot(surfaceaugm, hatbetam)
hatepsilonm = np.array(prix) - hatprixm
scr = np.dot(np.transpose(hatepsilonm), hatepsilonm)
```

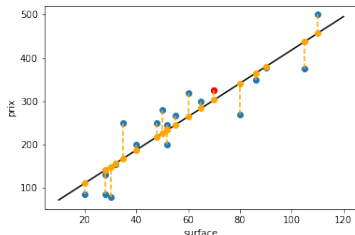
en noir : la droite $y = \hat{\beta}_0 + \hat{\beta}_1 x$;

points bleus : les données appartements initiales (x_i, y_i) ;

points oranges : les prédictions i.e. les couples (x_i, \hat{y}_i)

en orange pointillé : les longueurs des résidus

point rouge : le 14 ème appartement (70, 325).



ce qui donne

pour la somme des carrés des résidus : $scr \approx 36476.88$
pour la somme des carrés expliquée : $sce \approx 195068.321$
pour la somme des carrés totale : $sct \approx 231545.200$
pour le coefficient de détermination : $r^2 \approx 0.842$

Exemple pour $p = 2$ (basketteurs) :

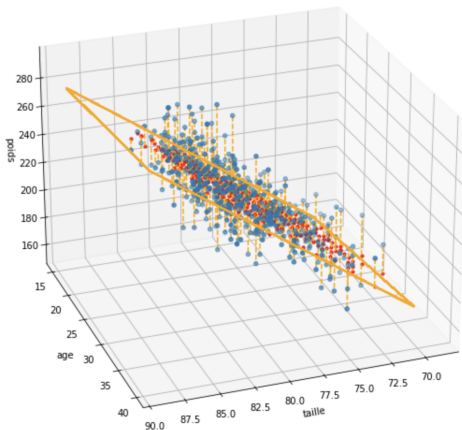
avec la fonction OLS :

```
hatprix = reg.fittedvalues  
hatepsilon = reg.resid  
scr = reg.ssr  
sce = reg.ess  
sct = reg.centered_tss  
r2 = reg.rsquared
```

ou encore, avec les notations matricielles :

```
hat_y = np.dot(x_tab, hat_beta)  
hat_epsilon = y_vect - hat_y  
scr = np.dot(np.transpose(hat_epsilon), hat_epsilon)
```

en rouge : le plan $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$;
points bleus : les données "joueurs de NBA" initiales
 (x_{i1}, x_{i2}, y_i) ;
points oranges : les prédictions i.e. les triplets $(x_{i1}, x_{i2}, \hat{y}_i)$
en orange pointillé : les longueurs des résidus



ce qui donne

pour la somme des carrés des résidus : $scr \approx 112784.883$

pour la somme des carrés expliquée : $sce \approx 244982.206$

pour la somme des carrés totale : $sct \approx 357767.089$

pour le coefficient de détermination : $r^2 \approx 0.685$

Vers la modélisation probabiliste des erreurs

Etant donné que nos points ne sont PAS sur un hyperplan affine, nous avons un terme d'erreur incompressible.

Nous allons modéliser ce terme d'erreur par un vecteur aléatoire.

Pourquoi modéliser de manière probabiliste ce terme d'erreur ?

Parce que ces erreurs peuvent provenir

- de causes inconnues, qui peuvent être en très grand nombre, et dont on n'a aucune idée,
- d'une part intrinsèque d'aléatoire dans la situation considérée

B) Modélisation probabiliste des erreurs

Nous allons donc à présent supposer que la **fluctuation** de y autour de l'hyperplan affine de régression linéaire théorique $y = \beta_0 + \sum_{j=1}^p \beta_j x_j$ (resp. $y = \beta_0 + \beta_1 x$ pour le cas de la RLS $p = 1$) est de nature **aléatoire** : autrement dit, nous supposons que y est une réalisation d'une v.a.r. Y où

$$Y = \beta_0 + \sum_{j=1}^p \beta_j x_j + \mathcal{E} \quad (\text{resp. } Y = \beta_0 + \beta_1 x + \mathcal{E} \text{ pour la RLS})$$

- Y est la variable à expliquer, à valeurs dans \mathbb{R}
- x_1, \dots, x_p sont les variables explicatives, chacune appartenant à \mathbb{R}
(resp. pr la RLS, x est la variable explicative, appartenant à \mathbb{R})
- \mathcal{E} est le terme d'**erreur aléatoire** du modèle, à valeurs dans \mathbb{R} ;
- $\beta_0, \beta_1, \dots, \beta_p$ sont les $(p + 1)$ paramètres inconnus à estimer, chacun appartenant à \mathbb{R}

(resp. pr la RLS, β_0 et β_1 sont les deux paramètres inconnus à estimer, tous deux appartenant à \mathbb{R})

Nous supposons donc que les données $((x_1, y_1), \dots, (x_n, y_n))$ (resp.

$((x_1, y_1), \dots, (x_n, y_n))$ pr la RLS) sont une réalisation de l'observation

$((X_1, Y_1), \dots, (X_n, Y_n))$ (resp. $((X_1, Y_1), \dots, (X_n, Y_n))$ pr la RLS) où :

- X_1, \dots, X_n vecteurs aléat. constants : $\forall i \in \llbracket 1, n \rrbracket, X_i = x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$
(resp. pr la RLS : X_1, \dots, X_n v.a.r. constantes : $\forall i \in \llbracket 1, n \rrbracket, X_i = x_i$)
- $\forall i \in \llbracket 1, n \rrbracket, Y_i = \beta_0 + \left\langle \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, x_i \right\rangle_{\mathbb{R}^p} + \mathcal{E}_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \mathcal{E}_i$
(resp. pr la RLS : $\forall i \in \llbracket 1, n \rrbracket, Y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i$)
- $\mathcal{E}_1, \dots, \mathcal{E}_n$ v.a.r. représentant les termes d'erreurs et vérifiant
 - (A1) "les erreurs sont centrées" : $\forall i \in \llbracket 1, n \rrbracket, \mathbb{E}[\mathcal{E}_i] = 0$
 - (A2) "condition d'**homoscédasticité**" : $\forall i \in \llbracket 1, n \rrbracket, \text{Var}[\mathcal{E}_i] = \sigma^2$
 - (A3) "les termes d'erreurs sont non corrélés" :
 $\forall i, i' \in \llbracket 1, n \rrbracket, i \neq i' \Rightarrow \text{Cov}(\mathcal{E}_i, \mathcal{E}_{i'}) = 0$
- $\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*$ (resp. $\beta_0, \beta_1 \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*$ pr la RLS) paramètres inconnus

L'observation est donc $((x_1, Y_1), \dots, (x_n, Y_n))$ (resp. $((x_1, Y_1), \dots, (x_n, Y_n))$ pr la RLS), avec les x_i déterministes ds \mathbb{R}^p (resp. x_i déterministes ds \mathbb{R} pr la RLS), et les Y_i aléatoires à valeurs ds \mathbb{R} .

On peut également réécrire matriciellement le modèle

$\forall i \in \llbracket 1, n \rrbracket, Y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \mathfrak{E}_i$ sous la forme $Y_{\cdot} = x_{\cdot} \beta_{\cdot} + \mathfrak{E}_{\cdot}$ où

- $Y_{\cdot} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathcal{M}_{n,1}(\mathbb{R}) \rightsquigarrow \begin{matrix} \text{vecteur à expliquer} \\ \text{(ALÉATOIRE \& OBSERVÉ)} \end{matrix},$
- $x_{\cdot} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \in \mathcal{M}_{n,(p+1)}(\mathbb{R}) \rightsquigarrow \begin{matrix} \text{matrice explicative} \\ \text{(DÉTERMINISTE \& OBSERVÉE)} \end{matrix},$
- $\beta_{\cdot} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathcal{M}_{(p+1),1}(\mathbb{R}) \rightsquigarrow \begin{matrix} \text{vecteur des paramètres} \\ \text{(DÉTERMINISTE \& INCONNU (NON OBSERVÉ))} \end{matrix},$
- $\mathfrak{E}_{\cdot} = \begin{pmatrix} \mathfrak{E}_1 \\ \vdots \\ \mathfrak{E}_n \end{pmatrix} \in \mathcal{M}_{n,1}(\mathbb{R}) \rightsquigarrow \begin{matrix} \text{vecteur d'erreurs} \\ \text{(ALÉATOIRE \& NON OBSERVÉ)} \end{matrix}.$

Les conditions (A1)-(A2)-(A3) se réécrivent matriciellement :

$$(A1) : \mathbb{E}[\mathfrak{E}_{\cdot}] = 0_{n,1} \quad \text{et} \quad (A2)-(A3) : \text{Var}[\mathfrak{E}_{\cdot}] = \sigma^2 I_n$$

(Ici on ne suppose PAS que les \mathfrak{E}_i sont indépendants, NI qu'ils sont de même loi, mais on le supposera ds le C))

Rappels : $\mathbb{E}[\mathbf{E}.] = \begin{pmatrix} \mathbb{E}[\mathbf{E}_1] \\ \vdots \\ \mathbb{E}[\mathbf{E}_n] \end{pmatrix}$ est le vecteur d'espérance du vecteur aléat. \mathbf{E} .

$\text{Var}[\mathbf{E}.]$ est la matrice de variance-covariance du vecteur aléat. \mathbf{E} , et est définie par $\text{Var}[\mathbf{E}.] = [\text{Cov}(\mathbf{E}_i, \mathbf{E}_k)]_{1 \leq i, k \leq n}$

$$= \begin{pmatrix} \text{Var}[\mathbf{E}_1] & \text{Cov}(\mathbf{E}_1, \mathbf{E}_2) & \dots & \text{Cov}(\mathbf{E}_1, \mathbf{E}_n) \\ \text{Cov}(\mathbf{E}_2, \mathbf{E}_1) & \text{Var}[\mathbf{E}_2] & & \vdots \\ \vdots & & \text{Var}[\mathbf{E}_{n-1}] & \text{Cov}(\mathbf{E}_{n-1}, \mathbf{E}_n) \\ \text{Cov}(\mathbf{E}_n, \mathbf{E}_1) & \dots & \text{Cov}(\mathbf{E}_n, \mathbf{E}_{n-1}) & \text{Var}[\mathbf{E}_n] \end{pmatrix}$$

Conséquences des conditions (A1)-(A2)-(A3) :

- $\mathbb{E}[Y_i] = x_i \beta$ c-à-d $\forall i \in \llbracket 1, n \rrbracket, \mathbb{E}[Y_i] = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$

Cela entraîne : $\mathbb{E}[\overline{Y}] = \beta_0 + \sum_{j=1}^p \beta_j \overline{x_{.j}}$

- $\text{Var}[Y_i] = \text{Cov}(Y_i, \mathbf{E}_i) = \sigma^2 I_n$;
c-à-d $\forall i, i' \in \llbracket 1, n \rrbracket, \rightsquigarrow$ si $i = i'$, $\text{Cov}(Y_i, Y_{i'}) = \text{Var}[Y_i] = \sigma^2$,
 \rightsquigarrow si $i \neq i'$, $\text{Cov}(Y_i, Y_{i'}) = \text{Cov}(Y_i, \mathbf{E}_{i'}) = 0$

Cela entraîne : $\forall i \in \llbracket 1, n \rrbracket, \text{Var}[\overline{Y}] = \text{Cov}(Y_i, \overline{Y}) = \sigma^2/n$

Exercice : prouvez ces égalités en justifiant chaque étape de calcul

Estimation des paramètres théoriques $\beta_0, \beta_1, \dots, \beta_p$

D'après le A), on construit à partir de l'observation

$((x_1., Y_1), \dots, (x_n., Y_n))$ les estimateurs aléatoires de $\beta_0, \beta_1, \dots, \beta_p$ suivants :

$$\hat{\beta}_{\cdot} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0((x_1., Y_1), \dots, (x_n., Y_n)) \\ \hat{\beta}_1((x_1., Y_1), \dots, (x_n., Y_n)) \\ \vdots \\ \hat{\beta}_p((x_1., Y_1), \dots, (x_n., Y_n)) \end{pmatrix} = ({}^t x_{..} x_{..})^{-1} {}^t x_{..} Y_{\cdot}$$

Pr une réalisation $((x_1., y_1), \dots, (x_n., y_n))$ de l'observation, on retrouve les valeurs

$$\hat{\beta}_{\cdot} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0((x_1., y_1), \dots, (x_n., y_n)) \\ \hat{\beta}_1((x_1., y_1), \dots, (x_n., y_n)) \\ \vdots \\ \hat{\beta}_p((x_1., y_1), \dots, (x_n., y_n)) \end{pmatrix} = ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y_{\cdot}$$

Pour la RLS, d'après le A), on construit à partir de l'observation $((x_1, Y_1), \dots, (x_n, Y_n))$ les estimateurs aléatoires de β_0, β_1 suivants :

$$\hat{\beta}_{\cdot} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0((x_1, Y_1), \dots, (x_n, Y_n)) \\ \hat{\beta}_1((x_1, Y_1), \dots, (x_n, Y_n)) \end{pmatrix} = \begin{pmatrix} \overline{Y_{\cdot}} - \hat{\beta}_1 \overline{x_{\cdot}} \\ \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x_{\cdot}})(Y_i - \overline{Y_{\cdot}})}{\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x_{\cdot}})^2} = \frac{\hat{c}_{x_{\cdot}, Y_{\cdot}}}{\hat{\sigma}_{x_{\cdot}}^2} = \frac{\hat{\sigma}_{Y_{\cdot}}}{\hat{\sigma}_{x_{\cdot}}} \hat{\rho}_{x_{\cdot}, Y_{\cdot}} \end{pmatrix}$$

et pr une réalisation $((x_1, y_1), \dots, (x_n, y_n))$ de l'observation, on retrouve les valeurs

$$\hat{\beta}_{\cdot} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0((x_1, y_1), \dots, (x_n, y_n)) \\ \hat{\beta}_1((x_1, y_1), \dots, (x_n, y_n)) \end{pmatrix} = \begin{pmatrix} \overline{y_{\cdot}} - \hat{\beta}_1 \overline{x_{\cdot}} \\ \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x_{\cdot}})(y_i - \overline{y_{\cdot}})}{\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x_{\cdot}})^2} = \frac{\hat{c}_{x_{\cdot}, y_{\cdot}}}{\hat{\sigma}_{x_{\cdot}}^2} = \frac{\hat{\sigma}_{y_{\cdot}}}{\hat{\sigma}_{x_{\cdot}}} \hat{\rho}_{x_{\cdot}, y_{\cdot}} \end{pmatrix}$$

Propriétés de l'estimateur $\hat{\beta}$ de β .

Propriétés de l'estimateur $\hat{\beta}$ de β . (A SAVOIR PAR COEUR)

Sous les conditions (A0)-(A1)-(A2)-(A3) : pour tous $(\beta, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_+^*$

- $\mathbb{E}_{(\beta, \sigma^2)}[\hat{\beta}] = \beta$. $\leadsto \hat{\beta}$ estimateur **sans biais** de β .
 $\forall j \in \llbracket 0, p \rrbracket$, $\mathbb{E}_{(\beta, \sigma^2)}[\hat{\beta}_j] = \beta_j$ $\leadsto \hat{\beta}_j$ estimateur **sans biais** de β_j

Rq : reste vrai pr la RLS avec $p = 1$

- $\text{Var}_{(\beta, \sigma^2)}[\hat{\beta}] = \sigma^2 ({}^t X_{..} X_{..})^{-1}$

Rq : pr la RLS, cela donne $\text{Var}_{(\beta, \sigma^2)}[\hat{\beta}] = \frac{\sigma^2}{n\hat{\sigma}_x^2} \begin{pmatrix} \hat{\sigma}_x^2 + (\bar{x})^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$

d'où $\text{Var}_{(\beta, \sigma^2)}[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{n\hat{\sigma}_x^2} \right)$, $\text{Var}_{(\beta, \sigma^2)}[\hat{\beta}_1] = \frac{\sigma^2}{n\hat{\sigma}_x^2}$ et $\text{Cov}_{(\beta, \sigma^2)}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{n\hat{\sigma}_x^2}$

Propriété de Gauss-Markov (ADMIS, pas à savoir) : $\hat{\beta}$ est de variance minimale parmi les estimateurs sans biais de β linéaires en Y_1, \dots, Y_n

Ds la suite, on omettra l'indice (β, σ^2) ; on reprendra cette notation plus loin pr les IC et les tests

Valeurs prédites & Résidus aléatoires : définition

- $\forall i \in \llbracket 1, n \rrbracket$, $\hat{Y}_i := x_i \cdot \hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}$ v.a.r. dite **valeur prédite** pour Y_i
 $\leadsto \hat{Y}_i = \hat{Y}_i((x_1., Y_1), \dots, (x_n., Y_n))$; pr une réal. $((x_1., y_1), \dots, (x_n., y_n))$ de l'obs. $((x_1., Y_1), \dots, (x_n., Y_n))$, $\hat{Y}_i((x_1., y_1), \dots, (x_n., y_n)) = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij} = \hat{y}_i$
- $\forall i \in \llbracket 1, n \rrbracket$, $\hat{\mathcal{E}}_i := Y_i - \hat{Y}_i$ v.a.r. dite **résidu** ds la prédiction de Y_i par \hat{Y}_i
 $\leadsto \hat{\mathcal{E}}_i = \hat{\mathcal{E}}_i((x_1., Y_1), \dots, (x_n., Y_n))$; pr une réal. $((x_1., y_1), \dots, (x_n., y_n))$ de l'obs. $((x_1., Y_1), \dots, (x_n., Y_n))$, $\hat{\mathcal{E}}_i((x_1., y_1), \dots, (x_n., y_n)) = y_i - \hat{y}_i = \hat{\epsilon}_i$
- $\min_{\beta. \in \mathbb{R}^{p+1}} \sum_{i=1}^n (Y_i - (\beta_0 + \sum_{j=1}^p \beta_j x_{ij}))^2 = \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}))^2 = \sum_{i=1}^n \hat{\mathcal{E}}_i^2$ v.a.r.
dite **somme des carrés des résidus (SCR)**

$\leadsto \text{SCR} = \text{SCR}((x_1., Y_1), \dots, (x_n., Y_n))$ et pr une réal.

$$((x_1., y_1), \dots, (x_n., y_n)), \text{SCR}((x_1., y_1), \dots, (x_n., y_n)) = \sum_{i=1}^n \hat{\epsilon}_i^2 = \text{scr}$$

Valeurs prédites & Résidus aléatoires : définition pour la RLS

- $\forall i \in \llbracket 1, n \rrbracket$, $\hat{Y}_i := x_i \cdot \hat{\beta}_\cdot = \hat{\beta}_0 + \hat{\beta}_1 x_i$ v.a.r. dite **valeur prédite** pour Y_i

$\leadsto \hat{Y}_i = \hat{Y}_i((x_1, Y_1), \dots, (x_n, Y_n))$; pr une réal. $((x_1, y_1), \dots, (x_n, y_n))$ de l'obs. $((x_1, Y_1), \dots, (x_n, Y_n))$,
 $\hat{Y}_i((x_1, y_1), \dots, (x_n, y_n)) = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{y}_i$

- $\forall i \in \llbracket 1, n \rrbracket$, $\hat{\epsilon}_i := Y_i - \hat{Y}_i$ v.a.r. dite **résidu** ds la prédiction de Y_i par \hat{Y}_i

$\leadsto \hat{\epsilon}_i = \hat{\epsilon}_i((x_1, Y_1), \dots, (x_n, Y_n))$; pr une réal. $((x_1, y_1), \dots, (x_n, y_n))$ de l'obs. $((x_1, Y_1), \dots, (x_n, Y_n))$,
 $\hat{\epsilon}_i((x_1, y_1), \dots, (x_n, y_n)) = y_i - \hat{y}_i = \hat{\epsilon}_i$

- $\min_{\beta_\cdot \in \mathbb{R}^{p+1}} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \sum_{i=1}^n \hat{\epsilon}_i^2$ v.a.r. dite **somme des carrés des résidus** (SCR)

$\leadsto \text{SCR} = \text{SCR}((x_1, Y_1), \dots, (x_n, Y_n))$ et pr une réal. $((x_1, y_1), \dots, (x_n, y_n))$, $\text{SCR}((x_1, y_1), \dots, (x_n, y_n)) = \sum_{i=1}^n \hat{\epsilon}_i^2 = \text{scr}$

Valeurs prédites & Résidus aléatoires : propriétés

Avec les notations matricielles $\hat{Y}_{\cdot} := \begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix}$, $\hat{\mathcal{E}}_{\cdot} := \begin{pmatrix} \hat{\mathcal{E}}_1 \\ \vdots \\ \hat{\mathcal{E}}_n \end{pmatrix}$, on a les propriétés :

- $\hat{Y}_{\cdot} = X_{\cdot\cdot} \hat{\beta}_{\cdot} = X_{\cdot\cdot} ({}^t X_{\cdot\cdot} X_{\cdot\cdot})^{-1} {}^t X_{\cdot\cdot} Y_{\cdot} = \Pi_{\text{Im}(X_{\cdot\cdot})}(Y_{\cdot})$
- $\mathbb{E}[\hat{Y}_{\cdot}] = X_{\cdot\cdot} \beta_{\cdot} = \mathbb{E}[Y_{\cdot}] \rightsquigarrow \hat{Y}_{\cdot}$ **estimateur sans biais** $\mathbb{E}[Y_{\cdot}]$
- $\forall i \in \llbracket 1, n \rrbracket$, $\mathbb{E}[\hat{Y}_i] = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} = \mathbb{E}[Y_i]$
 $\rightsquigarrow \hat{Y}_i$ **estimateur sans biais de** $\mathbb{E}[Y_i]$
- $\text{Var}[\hat{Y}_{\cdot}] = \sigma^2 X_{\cdot\cdot} ({}^t X_{\cdot\cdot} X_{\cdot\cdot})^{-1} {}^t X_{\cdot\cdot}$
- $\hat{\mathcal{E}}_{\cdot} = (I_n - X_{\cdot\cdot} ({}^t X_{\cdot\cdot} X_{\cdot\cdot})^{-1} {}^t X_{\cdot\cdot}) Y_{\cdot} = \Pi_{(\text{Im}(X_{\cdot\cdot}))^{\perp}}(Y_{\cdot})$
 $= (I_n - X_{\cdot\cdot} ({}^t X_{\cdot\cdot} X_{\cdot\cdot})^{-1} {}^t X_{\cdot\cdot}) \mathcal{E}_{\cdot} = \Pi_{(\text{Im}(X_{\cdot\cdot}))^{\perp}}(\mathcal{E}_{\cdot})$
- $\mathbb{E}[\hat{\mathcal{E}}_{\cdot}] = 0_{n,1}$ et $\text{Var}[\hat{\mathcal{E}}_{\cdot}] = \sigma^2 (I_n - X_{\cdot\cdot} ({}^t X_{\cdot\cdot} X_{\cdot\cdot})^{-1} {}^t X_{\cdot\cdot})$
- $\text{Cov}(\hat{Y}_{\cdot}, \hat{\mathcal{E}}_{\cdot}) = 0_{n,n}$

Exercice : prouvez ces égalités

Valeurs prédites & Résidus aléatoires : propriétés pour la RLS

- $\mathbb{E}[\hat{Y}_i] = \beta_0 + \beta_1 x_i = \mathbb{E}[Y_i] \rightsquigarrow \hat{Y}_i$ est un estimateur sans biais de $\mathbb{E}[Y_i]$;
 $\rightsquigarrow \mathbb{E}[\hat{\epsilon}_i] = 0$ car $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ par déf
- $\text{Var}[\hat{Y}_i] = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{n\hat{\sigma}_x^2} \right)$ et $\text{Var}[\hat{\epsilon}_i] = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{n\hat{\sigma}_x^2} \right)$
- $\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2\right] = \frac{n-2}{n} \sigma^2$
- $\forall i, i' \in \llbracket 1, n \rrbracket, \text{Cov}(\hat{Y}_i, \hat{\epsilon}_{i'}) = 0$

On peut également montrer les propriétés secondaires suivantes :

- $\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \hat{Y}_i\right] = \beta_0 + \beta_1 \bar{x} = \mathbb{E}[\overline{Y}]$ et $\text{Var}\left[\frac{1}{n} \sum_{i=1}^n \hat{Y}_i\right] = \frac{\sigma^2}{n}$
- $\text{Cov}(\hat{Y}_i, \hat{Y}_{i'}) = \text{Cov}(Y_i, \hat{Y}_{i'}) = \frac{\sigma^2}{n\hat{\sigma}_x^2} (\hat{\sigma}_x^2 + (x_i - \bar{x})(x_{i'} - \bar{x}))$
 $\rightsquigarrow \text{Cov}(\hat{\epsilon}_i, \hat{\epsilon}_{i'}) = \sigma^2 \delta_{ii'} - \frac{\sigma^2}{n\hat{\sigma}_x^2} (\hat{\sigma}_x^2 + (x_i - \bar{x})(x_{i'} - \bar{x}))$
 (où $\delta_{ii'} = 1$ si $i = i'$ et 0 sinon)

Rq : on a utilisé pour cela les propriétés intermédiaires suivantes :

- $\text{Cov}(Y_i, \hat{\beta}_0) = \frac{\sigma^2}{n\hat{\sigma}_x^2} (\hat{\sigma}_x^2 - \bar{x}(x_i - \bar{x}))$ et $\text{Cov}(\overline{Y}, \hat{\beta}_0) = \frac{\sigma^2}{n}$;
- $\text{Cov}(Y_i, \hat{\beta}_1) = \frac{\sigma^2}{n\hat{\sigma}_x^2} (x_i - \bar{x})$ et $\text{Cov}(\overline{Y}, \hat{\beta}_1) = 0$

Estimation du paramètre σ^2

Par déf, $\sigma^2 = \text{Var}[\mathcal{E}_i] = \text{Var}[Y_i] = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2]$. Comme \hat{Y}_i estimateur de $\mathbb{E}[Y_i]$, on estime σ^2 par une quantité proportionnelle à $\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$:

$$\begin{aligned}\hat{S}^2 = \hat{S}^2((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) &:= \frac{n}{n - (p + 1)} \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{n - (p + 1)} \sum_{i=1}^n \hat{\mathcal{E}}_i^2 = \frac{\text{SCR}}{n - (p + 1)}\end{aligned}$$

Pr une réal. $((x_{1.}, y_1), \dots, (x_{n.}, y_n))$ de l'obs.,

$$\hat{s}^2 := \hat{S}^2((x_{1.}, y_1), \dots, (x_{n.}, y_n)) = \frac{1}{n - (p + 1)} \sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{\text{scr}}{n - (p + 1)}$$

Cas de la RLS : $\hat{S}^2 = \hat{S}^2((x_1, Y_1), \dots, (x_n, Y_n)) := \frac{1}{n-2} \sum_{i=1}^n \hat{\mathcal{E}}_i^2 = \frac{\text{SCR}}{n-2}$ et pr une réal. $((x_1, y_1), \dots, (x_n, y_n))$ de l'obs.,

$$\hat{s}^2 := \hat{S}^2((x_1, y_1), \dots, (x_n, y_n)) = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{\text{scr}}{n-2}$$

$\mathbb{E}[\hat{S}^2] = \sigma^2 \quad \rightsquigarrow \hat{S}^2$ estimateur sans biais de σ^2

Rq : la perte de $(p + 1)$ degrés de liberté dans l'expression de \hat{S}^2 est le "coût" de l'estimation de β

Prévision de l'étiquette d'une nouvelle donnée

Supposons que l'on dispose d'une nouvelle donnée, non étiquetée, $(x_{n+1,1}, \dots, x_{n+1,p})$, mais qu'a priori l'on peut supposer son comportement "proche" de celui des données déjà connues.

Autrement dit, l'étiquette y_{n+1} de x_{n+1} doit vérifier $y_{n+1} \simeq \beta_0 + \sum_{j=1}^p \beta_j x_{n+1,j}$

On utilise alors comme prévision de y_{n+1} la valeur

$$\hat{y}_{n+1}^{\text{prev}} := \hat{\beta}_0 + \sum_{j=1}^p \beta_j x_{n+1,j} = x_{n+1,\cdot} \hat{\beta}. \quad \text{avec } x_{n+1,\cdot} := (1, x_{n+1,1}, \dots, x_{n+1,p}).$$

Si l'on suppose que y_{n+1} est, comme les étiquettes précédentes, une réalisation de $Y_{n+1} = \beta_0 + \sum_{j=1}^p \beta_j x_{n+1,j} + \mathfrak{E}_{n+1} = x_{n+1,\cdot} \beta + \mathfrak{E}_{n+1}$, on définit

$$\text{la prévision de } Y_{n+1} \text{ par } \hat{Y}_{n+1}^{\text{prev}} := \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{n+1,j} = x_{n+1,\cdot} \hat{\beta}.$$

et l'erreur dans la prévision de Y_{n+1} par $\hat{Y}_{n+1}^{\text{prev}}$ par $\hat{\mathfrak{E}}_{n+1}^{\text{prev}} := Y_{n+1} - \hat{Y}_{n+1}^{\text{prev}}$

Propriétés de $\hat{Y}_{n+1}^{\text{prev}}$ et $\hat{\epsilon}_{n+1}^{\text{prev}}$

- Si $\mathbb{E}[\epsilon_{n+1}] = 0$, alors $\mathbb{E}[\hat{Y}_{n+1}^{\text{prev}}] = \beta_0 + \sum_{j=1}^p \beta_j x_{n+1,j} = \mathbb{E}[Y_{n+1}]$
 $\leadsto \hat{Y}_{n+1}^{\text{prev}}$ estimateur sans biais de $\mathbb{E}[Y_{n+1}]$ (d'où $\mathbb{E}[\hat{\epsilon}_{n+1}^{\text{prev}}] = 0$)

Cas de la RLS : $\mathbb{E}[\hat{Y}_{n+1}^{\text{prev}}] = \beta_0 + \beta_1 x_{n+1} = \mathbb{E}[Y_{n+1}]$

- Si de plus $\text{cov}(\epsilon_{n+1}, \epsilon_{i'}) = \begin{cases} 0 & \text{si } i' \in \llbracket 1, n \rrbracket \\ \sigma^2 & \text{si } i' = n+1 \end{cases}$, alors

$$\text{Var}[\hat{Y}_{n+1}^{\text{prev}}] = \sigma^2 x_{n+1,.} ({}^t x_{..} x_{..})^{-1} {}^t x_{n+1,.}$$

$$\text{Cas de la RLS : } \text{Var}[\hat{Y}_{n+1}^{\text{prev}}] = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n \hat{\sigma}_x^2} \right)$$

Par conséquent, la variance augmente quand x_{n+1} s'éloigne du centre de gravité \bar{x} des étiquettes des points. Effectuer une prévision de y_{n+1} lorsque x_{n+1} est loin de \bar{x} est risqué, car dans cette situation, la variance de l'erreur de prévision peut être très grande ...

$$\text{Var}[\hat{\epsilon}_{n+1}^{\text{prev}}] = \sigma^2 + \text{Var}[\hat{Y}_{n+1}^{\text{prev}}]$$

$$\text{Cas de la RLS : } \text{Var}[\hat{\epsilon}_{n+1}^{\text{prev}}] = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n \hat{\sigma}_x^2} \right) \text{ dépend}$$

- de la variabilité intrinsèque σ^2 de $Y_{n+1} \leadsto$ cette source de variabilité ne peut pas être réduite
- de la variabilité due à l'imprécision des estimations de $\beta_0, \beta_1 \leadsto$ cette source de variabilité peut être réduite en augmentant le nombre n de données.

En supposant que les erreurs entre le modèle théorique et nos données sont centrées, de même variance et deux à deux non corrélées, on a déterminé l'espérance et la variance des prédictions et des prévisions.

On a donc obtenu des informations supplémentaires par rapport à la partie A) où l'on obtenait juste des valeurs numériques ponctuelles pour les prédictions et les prévisions.

Comment faire pour quantifier encore plus précisément le risque pris en déviant légèrement de ces valeurs numériques ?

Nous avons besoin de leurs lois, et pour cela également de la loi de $\hat{\beta}$ et de $\hat{\sigma}^2$!

C) Modèle linéaire gaussien

Nous nous plaçons enfin sous les conditions (A0) et

$$(A4) : \mathbf{\epsilon}_. \sim \mathcal{N}_n(0_{n,1}, \sigma^2 I_n), \text{ ou encore } \epsilon_1, \dots, \epsilon_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

qui équivaut à

$$Y_. \sim \mathcal{N}_n(x_.\beta_., \sigma^2 I_n)$$

ou encore à

$$Y_1, \dots, Y_n \text{ indépendants, avec } \forall i \in \llbracket 1, n \rrbracket, Y_i \sim \mathcal{N}(\beta_0 + \sum_{j=1}^p \beta_j x_{ij}, \sigma^2)$$

Cas de la RLS : cette condition équivaut à Y_1, \dots, Y_n indépendants, avec $\forall i \in \llbracket 1, n \rrbracket, Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$

Remarque : (A4) implique (A1)-(A2)-(A3)

Outil principal : le Théorème de Cochran

Théorème de Cochran

Soit $m \in \mathbb{N}^*$, \mathbb{R}^m muni du produit scalaire euclidien usuel $\langle \cdot, \cdot \rangle_{\mathbb{R}^m}$ et de la

norme associée $\|\cdot\|_{\mathbb{R}^m}$, $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}$ vect. aléat. gaussien de \mathbb{R}^m tq

$Z \sim \mathcal{N}_m(0_{m,1}, I_m)$ autrement dit $Z_1, \dots, Z_m \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$,

G un \mathbb{R} -s.e.v. de \mathbb{R}^m de dimension d , G^\perp l'orthogonal de G dans \mathbb{R}^m (on a donc $G \oplus G^\perp = \mathbb{R}^m$ et G^\perp de dimension $m - d$), et Π_G (resp Π_{G^\perp}) la matrice de la projection orthogonale sur G (resp G^\perp) dans la base canonique de \mathbb{R}^m .

Alors :

- $\Pi_G Z \sim \mathcal{N}_m(0_{m,1}, \Pi_G)$, et $\|\Pi_G Z\|_{\mathbb{R}^m}^2 \sim \chi^2(d)$
- $\Pi_{G^\perp} Z \sim \mathcal{N}_m(0_{m,1}, \Pi_{G^\perp})$, et $\|\Pi_{G^\perp} Z\|_{\mathbb{R}^m}^2 \sim \chi^2(m - d)$
- $\Pi_G Z$ et $\Pi_{G^\perp} Z$ sont indépendants (et donc $\|\Pi_G Z\|_{\mathbb{R}^m}^2$ et $\|\Pi_{G^\perp} Z\|_{\mathbb{R}^m}^2$ sont également indépendants).

C) 1. IC & Test de significativité sur $\hat{\beta}_j$, $j \in \llbracket 0, p \rrbracket$

Propriétés supplémentaires de $\hat{\beta}_\cdot, \hat{S}^2$

Sous les conditions (A0) & (A4),

- $\hat{\beta}_\cdot \sim \mathcal{N}_{p+1}(\beta_\cdot, \sigma^2(t_{X_\cdot X_\cdot})^{-1})$
- $\forall j \in \llbracket 0, p \rrbracket, \hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2((t_{X_\cdot X_\cdot})^{-1})_{(j+1),(j+1)})$;
- $\frac{(n-(p+1))\hat{S}^2}{\sigma^2} \sim \chi^2(n - (p + 1))$
- $\hat{\beta}_\cdot \perp\!\!\!\perp \frac{(n-(p+1))\hat{S}^2}{\sigma^2}$

$$\bullet \quad \forall j \in \llbracket 0, p \rrbracket, \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2((t_{X_\cdot X_\cdot})^{-1})_{(j+1),(j+1)}}}}{\sqrt{\frac{(n-(p+1))\hat{S}^2}{\sigma^2} / (n-(p+1))}} = \frac{\hat{\beta}_j - \beta_j}{\hat{S} \sqrt{((t_{X_\cdot X_\cdot})^{-1})_{(j+1),(j+1)}}} \sim t(n - (p + 1))$$

Sous la condition (A4),

- $\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{(\bar{x})^2}{n\hat{\sigma}_x^2}\right)\right)$ et $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{n\hat{\sigma}_x^2}\right)$
- $\frac{(n-2)\hat{S}^2}{\sigma^2} \sim \chi^2(n-2)$
- $(\hat{\beta}_0, \hat{\beta}_1) \perp \hat{S}^2$
- $\frac{\hat{\beta}_0 - \beta_0}{\hat{S} \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{n\hat{\sigma}_x^2}}} \sim t(n-2)$ ($\hat{S} \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{n\hat{\sigma}_x^2}}$ est un estimateur de $\sqrt{\text{Var}[\hat{\beta}_0]}$)
- $\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{S}}{\sqrt{n\hat{\sigma}_x^2}}} \sim t(n-2)$ ($\frac{\hat{S}}{\sqrt{n\hat{\sigma}_x^2}}$ est un estimateur de $\sqrt{\text{Var}[\hat{\beta}_1]}$)

Application numérique pour l'exemple des appartements : $\frac{\hat{S}}{\sqrt{n\hat{\sigma}_x^2}} \simeq 0.392$

```
s2 = scr / (n-2)
estimateursdhatbeta1 = np.sqrt( s2 / ( n * np.var(surface) ) )
```

ou bien, à l'aide de la fonction OLS :

```
estimateursdhatbeta = reg.bse[1]
matvar = scipy.linalg.inv(np.dot(np.transpose(surfaceaugm), surfaceaugm))
estimateursdhatbeta1 = np.sqrt(s2 * np.diag(matvar)[1])
```

Intervalle de confiance pour β_j , $j \in \llbracket 0, p \rrbracket$

Intervalle de confiance de niveau exact $1 - \alpha$ de β_j :

$$IC_{1-\alpha}^{(\beta_j)}((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) := \left[\hat{\beta}_j \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{(({}^tX..X..) ^{-1})_{(j+1),(j+1)}} \right]$$

$$\leadsto \forall (\beta_., \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_+^*, \mathbb{P}_{(\beta_., \sigma^2)} \left(\beta_j \in IC_{1-\alpha}^{(\beta_j)}((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) \right) = 1 - \alpha$$

\leadsto pour une réal. $((x_{1.}, y_1), \dots, (x_{n.}, y_n))$ de l'obs. $((x_{1.}, Y_1), \dots, (x_{n.}, Y_n))$,

$$\begin{aligned} ic_{1-\alpha}^{(\beta_j)} &:= IC_{1-\alpha}^{(\beta_j)}((x_{1.}, y_1), \dots, (x_{n.}, y_n)) \\ &= \left[\hat{\beta}_j \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{(({}^tX..X..) ^{-1})_{(j+1),(j+1)}} \right] \end{aligned}$$

$$\text{Cas de la RLS : } IC_{1-\alpha}^{(\beta_1)}((x_1, Y_1), \dots, (x_n, Y_n)) := [\hat{\beta}_1 \pm q_{1-\frac{\alpha}{2}}^{t(n-2)} \frac{\hat{S}}{\sqrt{n\hat{\sigma}_x^2}}]$$

Ex pr $p = 1$ (appartements) :

```
avec alpha = 5% ic_{0.95}^{(\beta_1)} \simeq [3.024, 4.672]
qalpha = scipy.stats.t.ppf(1-alpha/2, df = n-2)
precision = qalpha * np.sqrt(s2)
/np.sqrt(n*np.var(surface))
icmoinsalpha = hatbeta1 + np.array([-1,1]) * precision
```

Ex pr $p = 2$ (basketteurs) : IC de niveau $1 - \alpha = 95\%$
avec OLS : `reg.conf_int(alpha)`

| variable | borne inf | borne sup |
|-----------|-------------|-------------|
| taille | 5.946877 | 6.706443 |
| age | 0.347830 | 0.954364 |
| intercept | -327.648146 | -265.549080 |

Test de Student de significativité de β_j , $j \in \llbracket 0, p \rrbracket$

On procède au test bilatère de $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$

- Observation : $((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) \in (\mathbb{R}^p \times \mathbb{R})^n$ (noté (Y_1, \dots, Y_n))
 - $Y_{.} \sim \mathcal{N}(x_{..}\beta_{.}, \sigma^2 I_n)$ (rappel : $Y_{.} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $x_{..} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$, $\beta_{.} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$)
 - $\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+^*$ inconnus
- Statistique utilisée $T_n = T_n(Y_1, \dots, Y_n) := \frac{\hat{\beta}_j}{\hat{S} \sqrt{(({}^t x_{..} x_{..})^{-1})_{(j+1), (j+1)}}}$
 - sous $H_0 : \beta_j = 0$, $T_n \sim t(n - (p + 1))$
 - sous $H_1 : \beta_j \neq 0$, $|T_n|$ aura tendance à prendre des valeurs plus grandes que sous H_0

- Test exact de taille $\alpha \in]0, 1[: \varphi_\alpha(Y_1, \dots, Y_n) := \mathbb{1}_{|T_n| > q_{1-\frac{\alpha}{2}}^{t(n-(p+1))}}$

On a $\varphi_\alpha(Y_1, \dots, Y_n) = \mathbb{1}_{0 \in IC_{1-\alpha}^{(\beta_j)}((x_1, Y_1), \dots, (x_n, Y_n))}$

- p-valeur associée à une réalisation $(y_1, \dots, y_n) :$

$$\begin{aligned}\pi(y_1, \dots, y_n) &= \mathbb{P}_{\beta_j=0}(|T_n(Y_1, \dots, Y_n)| \geq |T_n(y_1, \dots, y_n)|) \\ &= 2(1 - F_{t(n-(p+1))}(|T_n(y_1, \dots, y_n)|))\end{aligned}$$

Remarques :

- rejeter H_0 signifie que β_j est significativement non nul ;
- β_j s'interprète comme le taux d'accroissement moyen de y en fonction d'une variation de x_j à $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p$ fixés

Exemple pour $p = 2$ (basketteurs) : avec OLS, on trouve

| | reg.tvalues | reg.pvalues |
|-------------------|-------------|---------------|
| variable | statistique | p-valeur |
| $j = 1$ taille | 32.729203 | 1.312927e-126 |
| $j = 2$ age | 4.218099 | 2.923456e-05 |
| $j = 0$ intercept | -18.767675 | 6.229313e-60 |

donc, pr $j = 0, 1, 2$, on rejette H_0 au niveau $\alpha = 5\%$ pr le test de significativité de β_j

Rq : si on ne regarde que les 50 premiers basketteurs, les p-valeurs sont resp. 0.461129, 0.132534, 0.354148 et donc, pr $j = 0, 1, 2$, on conserve H_0 au niveau $\alpha = 5\%$ pr le test de significativité de β_j

Cas de la RLS : Test de Student de significativité de β_1

Remarque : cela revient à tester l'absence de liaison linéaire entre les x_i et les y_i , puisque celle-ci se traduit par la nullité du coefficient directeur β_1 .

On procède donc au test bilatère de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$

- Observation : $((x_1, Y_1), \dots, (x_n, Y_n)) \in \mathbb{R}^{2n}$ (notation abrégée (Y_1, \dots, Y_n))

- $\forall i \in \llbracket 1, n \rrbracket, Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ et $Y_1, \dots, Y_n \perp\!\!\!\perp$

- $\beta_0, \beta_1 \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*$ inconnus

- Statistique utilisée $T(Y_1, \dots, Y_n) := \frac{\hat{\beta}_1}{\frac{\hat{s}}{\sqrt{n\hat{\sigma}_x^2}}}$

- sous $H_0 : \beta_1 = 0, T(Y_1, \dots, Y_n) \sim t(n-2)$

- sous $H_1 : \beta_1 \neq 0, |T(Y_1, \dots, Y_n)|$ aura tendance à prendre des valeurs plus grandes que sous H_0

- Test statistique exact, dit de Student, de taille $\alpha \in]0, 1[$: $\varphi_\alpha(Y_1, \dots, Y_n) := \mathbb{1}_{|T(Y_1, \dots, Y_n)| > q_{1-\frac{\alpha}{2}}^{t(n-2)}}$

Application numérique dans notre exemple : $\text{scr} \simeq 36476.879$ et $\hat{s}^2 \simeq 2026.493$, d'où $T(y_1, \dots, y_n) \simeq 9.81$;

au niveau $\alpha = 5\%$, on a donc $\varphi_{0.05}(y_1, \dots, y_n) = 0$ car $T(y_1, \dots, y_n) \simeq 9.81 > q_{0.975}^{t(18)} \simeq 2.101$

et l'on rejette donc $H_0 : \beta_1 = 0$ au niveau 5%

- p -valeur associée à une réalisation (y_1, \dots, y_n) :

$$\pi(y_1, \dots, y_n) = \mathbb{P}_{\beta_1=0}(|T(Y_1, \dots, Y_n)| \geq |T(y_1, \dots, y_n)|) = 2(1 - F_{t(n-2)}(|T(y_1, \dots, y_n)|))$$

Application numérique dans notre exemple : $\pi(y_1, \dots, y_n) \simeq 2(1 - F_{t(18)}(|9.81|)) \simeq 1.197e-08$ qui est très petit, la

réalisation de $T(Y_1, \dots, Y_n)$ que l'on a obtenue serait donc très aberrante sous H_0 , et l'on est donc très confiant dans

notre décision de rejeter H_0 .

```

n = len(surface)
statStudent = hatbeta1 / ( np.sqrt(s2) / ( np.sqrt(n) * np.std(surface) ) )
alpha = 0.05
quantilealpha = scipy.stats.t.ppf(1 - alpha/2, df = n-2)
phialpha = ( np.abs(statStudent) > quantilealpha )
if phialpha :
    print("On rejette H0 au niveau ", 100*alpha, "
else :
    print("On conserve H0 au niveau ", 100*alpha, "
pvalStudent = 2 * ( 1 - scipy.stats.t.cdf(np.abs(statStudent), df = n-2) )

```

ou bien, avec la fonction OLS

```

statStudent = reg.tvalues[1]
pvalStudent = reg.pvalues[1]

```

C)2. IP de Y_{n+1} & IC de $\mathbb{E}[Y_{n+1}]$

Pr $(x_{n+1,1}, \dots, x_{n+1,p})$ nvelle donnée non étiquetée, d'étiquette modélisée par $Y_{n+1} = \beta_0 + \sum_{j=1}^p \beta_j x_{n+1,j} + \mathfrak{E}_{n+1}$, où \mathfrak{E}_{n+1} centrée et décorrélée de $\mathfrak{E}_1, \dots, \mathfrak{E}_n$, de prévision $\hat{Y}_{n+1}^{\text{prev}} := \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{n+1,j}$ on a déterminé $\mathbb{E}[\hat{Y}_{n+1}^{\text{prev}}]$ et $\text{Var}[\hat{Y}_{n+1}^{\text{prev}}]$.

Si $\mathfrak{E}_1, \dots, \mathfrak{E}_n, \mathfrak{E}_{n+1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, avec $x_{n+1,\cdot} := (1, x_{n+1,1}, \dots, x_{n+1,p})$, on a :

- $\hat{Y}_{n+1}^{\text{prev}} \sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^p \beta_j x_{n+1,j} = x_{n+1,\cdot} \beta, \sigma^2 x_{n+1,\cdot} (t_{X..X..})^{-1} t_{X_{n+1,\cdot}} \right)$

Cas de la RLS : avec $x_{n+1} \in \mathbb{R}$ la nvelle donnée non étiquetée, $\hat{Y}_{n+1}^{\text{prev}} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{n+1}, \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n \hat{\sigma}_x^2} \right) \right)$

- $\frac{\hat{Y}_{n+1}^{\text{prev}} - x_{n+1,\cdot} \beta}{\sigma \sqrt{x_{n+1,\cdot} (t_{X..X..})^{-1} t_{X_{n+1,\cdot}}}} \sim \mathcal{N}(0, 1)$ (Cas de la RLS : $\frac{\hat{Y}_{n+1}^{\text{prev}} - (\beta_0 + \beta_1 x_{n+1})}{\sigma \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n \hat{\sigma}_x^2}}} \sim \mathcal{N}(0, 1)$)

- $\frac{\hat{Y}_{n+1}^{\text{prev}} - x_{n+1,\cdot} \beta}{\hat{S} \sqrt{x_{n+1,\cdot} (t_{X..X..})^{-1} t_{X_{n+1,\cdot}}}} \sim t(n - (p + 1))$ (Cas de la RLS : $\frac{\hat{Y}_{n+1}^{\text{prev}} - (\beta_0 + \beta_1 x_{n+1})}{\hat{S} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n \hat{\sigma}_x^2}}} \sim t(n - 2)$)

L'erreur de prévision est $\hat{\epsilon}_{n+1}^{\text{prev}} := Y_{n+1} - \hat{Y}_{n+1}^{\text{prev}}$. On a :

- $\hat{\epsilon}_{n+1}^{\text{prev}} \sim \mathcal{N}\left(0, \sigma^2 \left(1 + x_{n+1,.} ({}^t x_{..} x_{..})^{-1} {}^t x_{n+1,.}\right)\right)$

(Cas de la RLS : $\hat{\epsilon}_{n+1}^{\text{prev}} \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\hat{\sigma}_x^2}\right)\right)$)

- $\frac{\hat{\epsilon}_{n+1}^{\text{prev}}}{\sigma \sqrt{1 + x_{n+1,.} ({}^t x_{..} x_{..})^{-1} {}^t x_{n+1,.}}} \sim \mathcal{N}(0, 1)$

(Cas de la RLS : $\frac{\hat{\epsilon}_{n+1}^{\text{prev}}}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\hat{\sigma}_x^2}}} \sim \mathcal{N}(0, 1)$)

- $\frac{\hat{\epsilon}_{n+1}^{\text{prev}}}{\hat{S} \sqrt{1 + x_{n+1,.} ({}^t x_{..} x_{..})^{-1} {}^t x_{n+1,.}}} \sim t(n - (p + 1))$

(Cas de la RLS : $\frac{\hat{\epsilon}_{n+1}^{\text{prev}}}{\hat{S} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\hat{\sigma}_x^2}}} \sim t(n - 2)$)

Intervalle de prévision de Y_{n+1}

L'intervalle de prévision de Y_{n+1} au niveau $1 - \alpha$ est

$$IP_{1-\alpha}^{(Y_{n+1})}((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) \\ := \left[\hat{Y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{1 + x_{n+1,.} ({}^tX_{..}X_{..})^{-1} {}^tX_{n+1,.}} \right]$$

$$\leadsto \mathbb{P}\left(Y_{n+1} \in IP_{1-\alpha}^{(Y_{n+1})}((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) \right) = 1 - \alpha$$

\leadsto pr une réal. $((x_{1.}, y_1), \dots, (x_{n.}, y_n))$ de l'obs. $((x_{1.}, Y_1), \dots, (x_{n.}, Y_n))$,

$$ip_{1-\alpha}^{(Y_{n+1})} := IP_{1-\alpha}^{(Y_{n+1})}((x_{1.}, y_1), \dots, (x_{n.}, y_n)) \\ = \left[\hat{y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{1 + x_{n+1,.} ({}^tX_{..}X_{..})^{-1} {}^tX_{n+1,.}} \right]$$

RLS : $IP_{1-\alpha}^{(Y_{n+1})}((x_1, Y_1), \dots, (x_n, Y_n)) := \left[\hat{Y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-2)} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\hat{\sigma}_x^2}} \right]$ et pr une réal. $((x_1, Y_1), \dots, (x_n, Y_n))$

de l'obs. $((x_1, Y_1), \dots, (x_n, Y_n))$, $ip_{1-\alpha}^{(Y_{n+1})} := IP_{1-\alpha}^{(Y_{n+1})}((x_1, Y_1), \dots, (x_n, Y_n)) = \left[\hat{y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-2)} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\hat{\sigma}_x^2}} \right]$

Application numérique pour notre exemple des appartements : pour $x_{n+1} = 77$, $\hat{y}_{n+1}^{\text{prev}} \simeq 329.926$, et pour $\alpha = 5\%$,

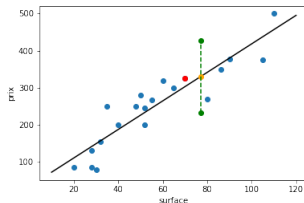
$ip_{0.95}^{(Y_{n+1})} \simeq [231.595, 428.257]$

```
xnplus1 = 77
alpha = 0.05
hatprevynplus1 = hatbeta0 + hatbeta1 * xnplus1
qalpha = scipy.stats.t.ppf(1 - alpha/2, df=n-2)
precision = qalpha * np.sqrt(s2 * ( 1 + 1/n + ( xnplus1 - np.mean(surface) )**2 / (n * np.var(surface))))
intervalleprev = hatprevynplus1 + np.array([-1,1]) * precision
```

ou bien, avec la fonction OLS :

```
xnplus1 = 77
xnplus1augm = np.array([1, xnplus1])
hatprevynplus1 = reg.predict(xnplus1augm)

regnou = reg.get_prediction(xnplus1augm)
intervalleprev = regnou.conf_int
                (obs=True, alpha=0.05)
```



On obtient une "région de prévision" pour y autour de l'hyperplan affine de régression linéaire estimé $(x_1, \dots, x_p) \mapsto \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_j$:

$$(x_1, \dots, x_p) \mapsto \left[\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_j \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{1 + x_{n+1,\cdot} (tX_{\cdot\cdot}X_{\cdot\cdot})^{-1} tX_{n+1,\cdot}} \right]$$

Cas de la RLS : on obtient une "région de prévision" pour y autour de la droite de régression linéaire estimée $x \mapsto \hat{\beta}_0 + \hat{\beta}_1 x$:

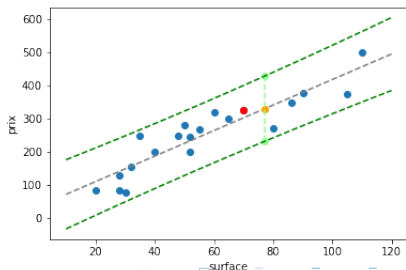
$$x \mapsto \left[\hat{\beta}_0 + \hat{\beta}_1 x \pm q_{1-\frac{\alpha}{2}}^{t(n-2)} \hat{S} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{n\hat{\sigma}_x^2}} \right]$$

```
xnplus = np.linspace(10, 120, num=1000)
```

```
alpha = 0.05
```

avec la fonction OLS :

```
xnplusaugm = statsmodels.api.add_constant(xnplus)
hatprevynplus = reg.predict(xnplusaugm)
regnouv = reg.get_prediction(xnplusaugm)
regionprevisiynplus = regnouv.conf_int(obs=True,
alpha=0.05)
borneinfprev = regionprevisiynplus[:,0]
bornesupprev = regionprevisiynplus[:,1]
```



ou bien :

```
hatprevynplus = hatbeta0 + hatbeta1 * xnplus
qalpha = scipy.stats.t.ppf(1 - alpha/2, df=n-2)
precision = qalpha * np.sqrt(s2 * (1 + 1/n + (xnplus-np.mean(surface))**2 / (n * np.var(surface))))
borneinfprev = hatprevynplus - precision
bornesupprev = hatprevynplus + precision
```

ou encore, avec les notations matricielles :

```
tsurfaceaugmm = np.transpose(surfaceaugmm)
approxvarhatbetam = s2 * scipy.linalg.inv(np.dot(tsurfaceaugmm, surfaceaugmm))
nnouv = len(xnplus)
xnplusaugmm = np.ones((nnouv, 2))
xnplusaugmm[:,1] = xnplus
hatprevynplum = np.dot(xnplusaugmm, hatbetam)
txnplusaugmm = np.transpose(xnplusaugmm)
approxvarprevm = np.dot(np.dot(xnplusaugmm, approxvarhatbetam), txnplusaugmm) + s2 * np.identity(nnouv)
precisionm = qalpha * np.sqrt(np.diag(approxvarprevm))
borneinfprevm = hatprevynplum - precisionm
bornesupprevm = hatprevynplum + precisionm
```

Intervalle de confiance de $\mathbb{E}[Y_{n+1}]$

Intervalle de confiance de $\mathbb{E}[Y_{n+1}] = \beta_0 + \sum_{j=1}^p \beta_j x_{n+1,j}$ de niveau $1 - \alpha$

$$IC_{1-\alpha}^{(\mathbb{E}[Y_{n+1}])}((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) \\ := \left[\hat{Y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{x_{n+1,.} ({}^t X_{..} X_{..})^{-1} {}^t x_{n+1,.}} \right]$$

$$\leadsto \mathbb{P}\left(\mathbb{E}[Y_{n+1}] \in IC_{1-\alpha}^{(\mathbb{E}[Y_{n+1}])}((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) \right) = 1 - \alpha$$

\leadsto pr une réal. $((x_{1.}, y_1), \dots, (x_{n.}, y_n))$ de l'obs. $((x_{1.}, Y_1), \dots, (x_{n.}, Y_n))$,

$$ic_{1-\alpha}^{(\mathbb{E}[Y_{n+1}])} := IC_{1-\alpha}^{(\mathbb{E}[Y_{n+1}])}((x_{1.}, y_1), \dots, (x_{n.}, y_n)) \\ = \left[\hat{y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{x_{n+1,.} ({}^t X_{..} X_{..})^{-1} {}^t x_{n+1,.}} \right]$$

RLS : Intervalle de confiance de $\mathbb{E}[Y_{n+1}] = \beta_0 + \beta_1 x_{n+1}$ de niveau $1 - \alpha$

$$IC_{1-\alpha}^{(\mathbb{E}[Y_{n+1}])}((x_1, Y_1), \dots, (x_n, Y_n)) := \left[\hat{Y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-2)} \hat{S} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\hat{\sigma}_x^2}} \right]$$

\leadsto pour une réalisation $((x_1, y_1), \dots, (x_n, y_n))$ de l'observation $((x_1, Y_1), \dots, (x_n, Y_n))$,

$$ic_{1-\alpha}^{(\mathbb{E}[Y_{n+1}])} := IC_{1-\alpha}^{(\mathbb{E}[Y_{n+1}])}((x_1, y_1), \dots, (x_n, y_n)) = \left[\hat{y}_{n+1}^{\text{prev}} \pm q_{1-\frac{\alpha}{2}}^{t(n-2)} \hat{s} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\hat{\sigma}_x^2}} \right]$$

Application numérique pour notre exemple des appartements : pour $x_{n+1} = 77$, $\hat{y}_{n+1}^{\text{prev}} \simeq 329.926$, et pour $\alpha = 5\%$,

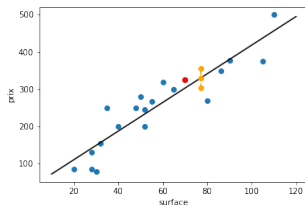
$$ic_{0.95}^{(\mathbb{E}[Y_{n+1}])} \simeq [303.014, 356.838]$$

```
xnplus1 = 77
alpha = 0.05
hatprevynplus1 = hatbeta0 + hatbeta1 * xnplus1
qalpha = scipy.stats.t.ppf(1 - alpha/2, df=n-2)
precision = qalpha * np.sqrt(s2 * ( 1/n + ( xnplus1 - np.mean(surface) )**2 / (n * np.var(surface))))
intervalleconf = hatprevynplus1 + np.array([-1,1]) * precision
```

ou bien, avec la fonction OLS :

```
xnplus1 = 77
xnplusiaugm = np.array([1, xnplus1])
hatprevynplus1 = reg.predict(xnplusiaugm)

regnouv = reg.get_prediction(xnplusiaugm)
intervalleconf = regnouv.conf_int
                (obs=False, alpha=0.05)
```



On obtient une "région de confiance" pour la valeur moyenne de y autour de l'hyperplan affine de régression linéaire estimé $(x_1, \dots, x_p) \mapsto \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_j$:

$$(x_1, \dots, x_p) \mapsto \left[\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_j \pm q_{1-\frac{\alpha}{2}}^{t(n-(p+1))} \hat{S} \sqrt{x_{n+1, \cdot} (t_{X \cdot \cdot} X \cdot \cdot)^{-1} t_{X_{n+1, \cdot}}} \right]$$

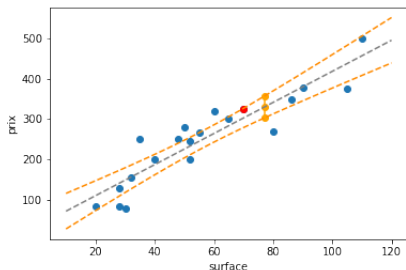
Cas de la RLS : On obtient une "région de confiance" pour la valeur moyenne de y autour de la droite de régression linéaire estimée $x \mapsto \hat{\beta}_0 + \hat{\beta}_1 x$ par

$$x \mapsto \left[\hat{\beta}_0 + \hat{\beta}_1 x \pm q_{1-\frac{\alpha}{2}}^{t(n-2)} \hat{S} \sqrt{\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{n \hat{\sigma}_x^2}} \right]$$

```
xnplus = np.linspace(10, 120, num=1000)
alpha = 0.05
```

avec la fonction OLS :

```
xnplusaugm = statsmodels.api.add_constant(xnplus)
hatprevynplus = reg.predict(xnplusaugm)
regnouv = reg.get_prediction(xnplusaugm)
regionconfianceespyplus = regnouv.conf_int(obs=False,
alpha=0.05)
borneinfconf = regionconfianceespyplus[:,0]
bornesupconf = regionconfianceespyplus[:,1]
```



ou bien :

```
hatprevynplus = hatbeta0 + hatbeta1 * xnplus
qalpha = scipy.stats.t.ppf(1 - alpha/2, df=n-2)
precision = qalpha * np.sqrt(s2 * ( 1/n + (xnplus-np.mean(surface))**2 / (n * np.var(surface))))
borneinfconf = hatprevynplus - precision
bornesupconf = hatprevynplus + precision
```

ou encore, avec les notations matricielles :

```
tsurfaceaugmm = np.transpose(surfaceaugmm)
approxvarhatbetam = s2 * scipy.linalg.inv(np.dot(tsurfaceaugmm, surfaceaugmm))
nnouv = len(xnplus)
xnplusaugmm = np.ones((nnouv, 2))
xnplusaugmm[:,1] = xnplus
hatprevynplum = np.dot(xnplusaugmm, hatbetam)
txnplusaugmm = np.transpose(xnplusaugmm)
approxvarconfm = np.dot(np.dot(xnplusaugmm, approxvarhatbetam), txnplusaugmm)
precisionm = qalpha * np.sqrt(np.diag(approxvarconfm))
borneinfconfm = hatprevynplum - precisionm
bornesupconfm = hatprevynplum + precisionm
```

C)3. Pertinence du modèle

Table d'ANOVA :

| Variabilité | Somme des carrés | ddl (degrés de liberté) |
|-------------------|--|------------------------------------|
| Expliquée | $SCE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ | p 1 pr la RLS |
| Résiduelle | $SCR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ | $n - (p + 1)$ $n - 2$ pr la RLS |
| Totale | $SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2$ | $n - 1$ |

$$\underbrace{\text{variabilité totale}}_{SCT} = \underbrace{\text{variabilité expliquée}}_{SCE} + \underbrace{\text{variabilité résiduelle}}_{SCR}$$

Le **coefficient de détermination** est la v.a.r.

$$R^2 = R^2(Y_1, \dots, Y_n) := \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Pr une réal. $((x_1, y_1), \dots, (x_n, y_n))$ de l'obs. $((x_1, Y_1), \dots, (x_n, Y_n))$, on a $r^2 = R^2(y_1, \dots, y_n) = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$

Test de Fisher issu de l'ANOVA

de l'absence de liaison linéaire entre les x_i et les y_i

L'absence de liaison linéaire entre les x_i et les y_i se traduit par la nullité de β_1, \dots, β_p . On procède donc au test bilatère :

$$H_0 : \beta_1 = \dots = \beta_p = 0 \text{ contre } H_1 : \exists j \in \llbracket 1, p \rrbracket, \beta_j \neq 0$$

- Observation : $((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)) \in (\mathbb{R}^p \times \mathbb{R})^n$ (noté (Y_1, \dots, Y_n))
 - $Y_{.} \sim \mathcal{N}(x_{.} \beta_{.}, \sigma^2 I_n)$ (rappel : $Y_{.} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $x_{.} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$, $\beta_{.} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$)
 - $\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+^*$ inconnus
- Statistique de Fisher : $T_n = T_n(Y_1, \dots, Y_n) := \frac{SCE/p}{SCR/(n-(p+1))}$
 - sous H_0 , $T_n \sim \mathcal{F}(p, n - (p + 1))$ (\mathcal{F} étant la loi de Fisher) .
 - sous H_1 , T_n aura tendance à prendre des valeurs plus grandes que sous H_0

- Test statistique exact, dit de Fisher, de taille $\alpha \in]0, 1[$:

$$\varphi_\alpha(Y_1, \dots, Y_n) = \mathbb{1}_{T_n(Y_1, \dots, Y_n) > q_{1-\alpha}^{\mathcal{F}(p, n-(p+1))}}$$

- p-valeur associée à une réalisation (y_1, \dots, y_n) :

$$\begin{aligned}\pi(y_1, \dots, y_n) &= \mathbb{P}_{\beta_1=0, \dots, \beta_p=0}(T_n(Y_1, \dots, Y_n) \geq T_n(y_1, \dots, y_n)) \\ &= 1 - F_{\mathcal{F}(p, n-(p+1))}(T_n(y_1, \dots, y_n))\end{aligned}$$

Ex pr $p = 2$ (basketteurs) : on se restreint aux $n = 50$ premiers, on trouve

$$T_{50}(y_1, \dots, y_{50}) \simeq 1.615$$

$$\text{pr } \alpha = 5\%, q_{1-\alpha}^{\mathcal{F}(p, n-(p+1))} = q_{0.95}^{\mathcal{F}(2, 47)} \simeq 3.195 \text{ donc}$$

$$T_{50}(y_1, \dots, y_{50}) \leq q_{1-\alpha}^{\mathcal{F}(p, n-(p+1))} \text{ et l'on conserve } H_0 \text{ au niveau } \alpha = 5\%$$

on trouve comme p-valeur : $\pi(y_1, \dots, y_{50}) \simeq 0.21$ donc

pour $\alpha < 0.21$ on conserve H_0 au niveau α

pour $\alpha > 0.21$ on rejette H_0 au niveau α et on accepte H_1

```
stat_Fisher = reg.fvalue
```

```
seuil_alpha = stats.f.ppf(1 - alpha, dfn=p, dfd = n-(p+1))
```

```
pval_Fisher = reg.f_pvalue
```

Cas de la RLS : Test statistique de Fisher issu de l'ANOVA de significativité du modèle, autrement dit de l'absence de liaison linéaire entre les x_i et les y_i , c-à-d test bilatère de $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$

- Observation : $((x_1, Y_1), \dots, (x_n, Y_n)) \in (\mathbb{R} \times \mathbb{R})^n$ (noté (Y_1, \dots, Y_n))
 - $Y_i \sim \mathcal{N}(x_i \beta_1, \sigma^2 I_n)$ (rappel : $Y_i = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $x_i = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\beta_i = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$)
 - $\beta_0, \beta_1 \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+^*$ inconnus
- Statistique utilisée : $T(Y_1, \dots, Y_n) := \frac{SCE}{\frac{SCR}{n-2}}$
 - sous H_0 , $T(Y_1, \dots, Y_n) \sim \mathcal{F}(1, n-2)$ (loi de Fisher) Rq : $\frac{SCE}{\frac{SCR}{n-2}} = \left(\frac{\hat{\beta}_1}{\hat{s}/\sqrt{n\hat{\sigma}_x^2}} \right)^2$ et, par définition de la loi de Fisher, le carré d'une v.a.r. de loi $t(n-2)$ est de loi $\mathcal{F}(1, n-2)$
 - sous H_1 , $T(Y_1, \dots, Y_n)$ aura tendance à prendre des valeurs plus grandes que sous H_0
- Test exact, dit de Fisher, de taille $\alpha \in]0, 1[$: $\varphi_\alpha(Y_1, \dots, Y_n) = \mathbb{1}_{T(Y_1, \dots, Y_n) > q_{1-\alpha}^{\mathcal{F}(1, n-2)}}$
 Application numérique dans notre exemple : $T(y_1, \dots, y_{20}) = \frac{sce}{\frac{scr}{n-2}} \simeq 96.259$, et pour $\alpha = 5\%$,
 $q_{1-\alpha}^{\mathcal{F}(1, n-2)} = q_{0.95}^{\mathcal{F}(1, 18)} \simeq 4.414$ donc on retrouve le fait que l'on rejette $H_0 : \beta_1 = 0$ au niveau 5%
- p-valeur associée à une réalisation (y_1, \dots, y_n) :
 $\pi(y_1, \dots, y_n) = \mathbb{P}_{\beta_1=0}(T(Y_1, \dots, Y_n) \geq T(y_1, \dots, y_n)) = 1 - F_{\mathcal{F}(1, n-2)}(T_n(y_1, \dots, y_n))$
 Application numérique dans notre exemple : $\pi(y_1, \dots, y_{20}) \simeq 1 - F_{\mathcal{F}(1, 18)}(96.259) \simeq 1.197e-08$

```
statFisher = sce/s2
quantilealpha = scipy.stats.f.ppf(1 - alpha, dfn=1, dfd = n-2)
phialpha = (statFisher > quantilealpha)
pvalFisher = 1 - scipy.stats.f.cdf(statFisher, dfn=1, dfd = n-2)
```

Remarque : ce test de Fisher est ÉGAL au test de Student de nullité de β_1 (attention, c'est spécifique à la RLS)

D) Limites du modèle & Sélection de variables

Nous allons enfin, à la lumière des résultats obtenus, procéder à un retour critique sur la qualité de notre modèle et sur les conditions que nous avons imposées initialement.

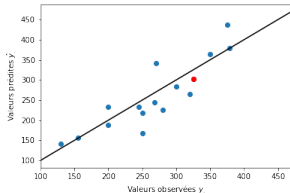
Objectifs :

- Etude de la qualité du modèle
- Retour critique sur les conditions (A0)-(A1)-(A2)-(A3)-(A4)
- Voir si nous ne pouvons pas supprimer certaines des p variables explicatives pour "alléger" le modèle
- Essayer de déterminer si notre modèle est suffisamment "robuste" par rapport aux données, en regardant quels sont les points leviers et s'il y a des outliers.

D)1. Adéquation du modèle

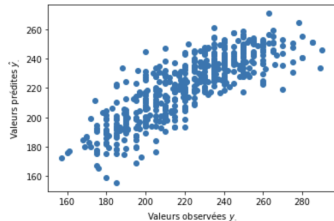
On peut évaluer graphiquement l'adéquation du modèle en traçant le graphique $(y_i, \hat{y}_i)_{i \in \llbracket 1, n \rrbracket}$ puisque les \hat{y}_i devaient "approximer" les y_i : les points doivent être alignés le long de la première bissectrice, cela signifie que l'adéquation du modèle aux données est correcte
→ si ce n'est pas le cas il y a une mauvaise adéquation du modèle aux données et il faut changer de modèle

Exemple pour $p = 1$ (appartements) :



→ les points sont bien proches de la 1ère bissectrice

Exemple pour $p = 2$ (basketteurs) :



→ les points sont plutôt proches de la 1ère bissectrice, mais pas les points extrêmes

D)2. Retour critique sur les conditions (A0-A4)

- Condition (A0) :

- $n > p + 1 \rightsquigarrow$ vérification immédiate
- ${}^t\mathbf{x}_{..}\mathbf{x}_{..}$ inversible \rightsquigarrow on vérifie que son déterminant est non nul :
`linalg.det(np.dot(np.transpose(x_tab), x_tab)) != 0`

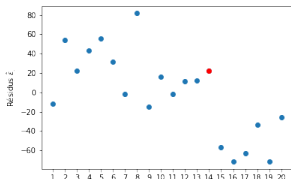
- Conditions (A1)-(A2)-(A3) : $\mathcal{E}_1, \dots, \mathcal{E}_n$ devaient être aléatoires, centrés, de même variance (cond. d'homoscédasticité) et non-corrélés entre eux
Pb : nous n'avons pas accès aux $\mathcal{E}_i \dots$

\rightsquigarrow mais nous avons accès aux résidus $\hat{\mathcal{E}}_i$: ils doivent avoir un
"comportement aléatoire, sans structure évidente" (pente, entonnoir, ...)

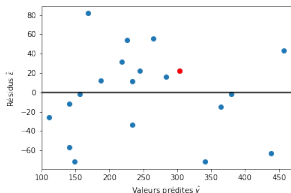
- on trace le graphe des $(i, \hat{\mathcal{E}}_i)_{i \in \llbracket 1, n \rrbracket} \rightsquigarrow$ si ce graphe présente une structure, alors une des conditions (A1)-(A2)-(A3) peut ne pas être vérifiée : il peut par exemple y avoir une autocorrélation
- on trace le graphes des $(\hat{y}_i, \hat{\mathcal{E}}_i)_{i \in \llbracket 1, n \rrbracket}$ et des $(\hat{y}_i, \hat{\mathcal{E}}_i^{(t \text{ ext})})_{i \in \llbracket 1, n \rrbracket}$
 \rightsquigarrow une structure particulière dans ces graphes peut indiquer que l'homoscédasticité n'est pas vérifiée

S'il y a une structure évidente, il faut changer de modèle pour essayer de prendre en compte cette structure (par exemple rajouter un terme quadratique x^2 ou polynomial dans les variables explicatives, ...)

Exemple pour $p = 1$



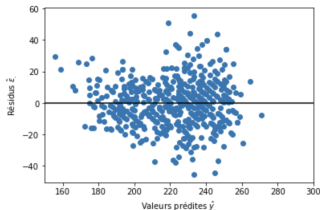
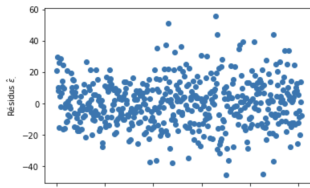
~ il semblerait que les $\hat{\epsilon}_i$
aient tendance à décroître...



~ alors qu'ici il n'y a pas
de structure particulière à signaler

Exemple pour $p = 2$

~ la variance semble augmenter avec i (les joueurs
sont classés par taille croissante)



~ structure en entonnoir, la variance augmente
avec les \hat{y}_i (elle dépend donc des x_i) : il y a
hétéroscédasticité et il faut changer de modèle pour
essayer de la prendre en compte

Notations supplémentaires

Pour $i \in \llbracket 1, n \rrbracket$, on note $h_{ii} = \left(\Pi_{\text{Im}(x_{..})} \right)_{ii} = \left(x_{..} ({}^t x_{..} x_{..})^{-1} {}^t x_{..} \right)_{ii}$ la i ème valeur sur la diagonale de la matrice de projection orthogonale sur $\text{Im}(x_{..})$ (= hat matrix) $\leadsto h_{ii}$ représente le poids de l'obs. i sur sa propre estimation

- i ème résidu normalisé : $\hat{\mathcal{E}}_i^{(1)} := \frac{\hat{\mathcal{E}}_i}{\sigma \sqrt{1-h_{ii}}}$ (rappel : $\text{Var}[\hat{\mathcal{E}}_i] = \sigma^2(1-h_{ii})$)
(attention, ils ne sont pas observables à cause de σ inconnu)
- i ème résidu standardisé : $\hat{\mathcal{E}}_i^{(t)} := \frac{\hat{\mathcal{E}}_i}{\hat{S} \sqrt{1-h_{ii}}}$ (Cas de la RLS : $\hat{\mathcal{E}}_i^{(t)} := \frac{\hat{\mathcal{E}}_i}{\hat{S} \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{n \hat{\sigma}_x^2}}}$)

Attention : $\hat{\mathcal{E}}_1^{(t)}, \dots, \hat{\mathcal{E}}_n^{(t)}$ ne sont pas indépendants et ne peuvent donc pas être représentatifs d'une absence/présence de structuration par autocorrélation.

- i ème résidu studentisé (externe) : $\hat{\mathcal{E}}_i^{(t \text{ ext})} := \hat{\mathcal{E}}_i^{(t)} \sqrt{\frac{n-(p+2)}{n-(p+1) - (\hat{\mathcal{E}}_i^{(t)})^2}}$
 $\leadsto \text{reg.outlier.test}()[:,0]$

Tous ces résidus sont de même variance 1.

- Condition (A4) : $\mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ pour un certain σ^2 c-à-d que nous avons supposé que les erreurs sont aléatoires, indépendantes, centrées, gaussiennes, de même variance.

Nous n'avons pas accès aux \mathcal{E}_i , mais nous avons accès aux réalisations des résidus $\hat{\mathcal{E}}_i$ et aux résidus studentisés $\hat{\mathcal{E}}_i^{(t \text{ ext})}$.

Or, si (A4) est vérifiée, et si retirer la i ème ligne de la matrice $x_{..}$ ne change pas son rang, $\hat{\mathcal{E}}_i^{(t \text{ ext})} \sim t(n - (p + 2))$

Nous allons donc vérifier sur ces résidus que la condition (A4) sur les $\hat{\mathcal{E}}_i$ n'est pas aberrante

Nous utilisons pour cela :

- des méthodes graphiques
- des tests

D)3. Sélection de variables

On dispose d'un modèle de RLM à p variables explicatives. On souhaite simplifier notre modèle et ne garder que certaines variables explicatives.

Comment les choisir au mieux ? Nous avons besoin :

- d'un **critère de qualité d'un modèle**, permettant de comparer deux modèles n'ayant pas nécessairement le même nombre de variables explicatives
- d'une **procédure de choix de modèles**, qui permet de choisir, parmi tous les modèles, le meilleur au sens de ce critère.

Problème de complexité : le nombre de modèles à considérer est

$\sum_{q=1}^p C_p^q = 2^p - 1$, qui croît exponentiellement avec p . Par exemple, si $p = 30$, on devrait considérer $2^{30} = 10^9$ modèles...

En pratique : on utilise des heuristiques dont les plus simples sont les procédures pas à pas ascendante/descendante

Critères de qualité d'un modèle

Le premier critère de qualité qui mesure l'ajustement du modèle aux données que nous avons vu est le **coefficient de détermination**

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\tilde{\sigma}_{\mathcal{E}}^2}{\tilde{\sigma}_Y^2}$$

Problème : le R^2 augmente lorsque le nombre de variables explicatives incluses dans le modèle augmente et il ne permet donc pas de comparer deux modèles n'ayant pas le même nombre de variables explicatives, seulement deux modèles ayant le même nombre de variables explicatives.

Pour pallier cette difficulté, on introduit le **coefficient de détermination ajusté**

$$R_{\text{ajusté}}^2 = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = 1 - \frac{\frac{n-1}{n-p-1} \tilde{\sigma}_{\mathcal{E}}^2}{\tilde{\sigma}_Y^2}$$

→ $R_{\text{ajusté}}^2$ n'augmente pas forcément lorsque le nombre de variables incluses dans le modèle augmente, et permet de comparer des modèles ayant un nombre de variables différent

Si notre modèle de RLM dispose d'une structure permettant de lui associer une vraisemblance, ce qui est le cas dans le cadre du C) Modèle Linéaire Gaussien, on dispose également d'un critère de vraisemblance pénalisé, le **Critère d'Information de Akaike** (Akaike Information Criterion) :

$$AIC := -2\ell_n^* + 2k$$

où

- ℓ_n^* est la log-vraisemblance maximisée
- k est le nombre de paramètres libres du modèle : ds le modèle de RLM à q variables explicatives, il y a $q + 2$ paramètres, $\beta_0, \beta_1, \dots, \beta_q, \sigma^2$, et une équation, $y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$, donc $k = (q + 2) - 1 = q + 1$ paramètres libres
- la fct de vraisemblance associée à une réal. $((x_{1.}, y_1), \dots, (x_{n.}, y_n))$ de l'obs. $((x_{1.}, Y_1), \dots, (x_{n.}, Y_n))$ est la densité jointe de $((x_{1.}, Y_1), \dots, (x_{n.}, Y_n))$ au point $((x_{1.}, y_1), \dots, (x_{n.}, y_n))$
 $L_n(\cdot; ((x_{1.}, y_1), \dots, (x_{n.}, y_n))) : (\beta., \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_+^*$

$$\mapsto f_{((x_{1.}, Y_1), \dots, (x_{n.}, Y_n)); (\beta., \sigma^2)}((x_{1.}, y_1), \dots, (x_{n.}, y_n)) = \frac{e^{-\frac{1}{2\sigma^2} \|y. - x. \cdot \beta.\|_{\mathbb{R}^n}^2}}{(2\pi\sigma^2)^{n/2}}$$

- l'estimateur du maximum de vraisemblance de (β, σ^2) est

$$(\hat{\beta}^{MV}, \hat{\sigma}^{2MV}) = \left(({}^t x_{..} x_{..})^{-1} {}^t x_{..} Y_{.}, \frac{SCR}{n} \right)$$

Preuve (même méthode que Ex 3 Feuille 1 de TD) : on commence par optimiser en β : pour $\sigma_0^2 \in \mathbb{R}_+^*$ fixé, $\beta \in \mathbb{R}^{p+1} \mapsto \ell_n(\beta, \sigma_0^2; ((x_{1.}, y_1), \dots, (x_{n.}, y_n))) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma_0^2) - \frac{1}{2\sigma_0^2} \|y_{.} - x_{..} \beta\|_{\mathbb{R}^n}^2$ est deux fois

dérivable et concave, de plus $\nabla_{\beta} \ell_n(\beta, \sigma_0^2; ((x_{1.}, y_1), \dots, (x_{n.}, y_n))) = \frac{1}{\sigma_0^2} {}^t x_{..} (y_{.} - x_{..} \beta)$, donc

$\ell_n(\cdot, \sigma_0^2; ((x_{1.}, y_1), \dots, (x_{n.}, y_n)))$ atteint son maximum en son unique point critique $\hat{\beta} = ({}^t x_{..} x_{..})^{-1} {}^t x_{..} y_{.}$.

On optimise maintenant la fonction $\sigma^2 \in \mathbb{R}_+^* \mapsto \ell_n(\hat{\beta}, \sigma^2; ((x_{1.}, y_1), \dots, (x_{n.}, y_n))) =$

$$-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \|y_{.} - x_{..} \hat{\beta}\|_{\mathbb{R}^n}^2 = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} scr$$

On a $\partial_{\sigma^2} \ell_n(\hat{\beta}, \sigma^2; ((x_{1.}, y_1), \dots, (x_{n.}, y_n))) = -\frac{n}{2\sigma^2} + \frac{scr}{2(\sigma^2)^2} = \frac{1}{2(\sigma^2)^2} (-n\sigma^2 + scr)$ qui est ≥ 0 ssi $\sigma^2 \leq \frac{scr}{n}$

donc $\ell_n(\hat{\beta}, \cdot; ((x_{1.}, y_1), \dots, (x_{n.}, y_n)))$ atteint son maximum en $\hat{\sigma}^2 = \frac{scr}{n}$

- le maximum de la log-vraisemblance est :

$$\ell_n^* = \ell_n(\hat{\beta}^{MV}, \hat{\sigma}^{2MV}; ((x_{1.}, y_1), \dots, (x_{n.}, y_n))) = -\frac{n}{2} (\ln(2\pi \hat{\sigma}^{2MV}) + 1)$$

On obtient donc

$$AIC = n \ln(SCR) + 2k + cste$$

Ces critères, du coefficient de détermination ajusté $R^2_{\text{ajusté}}$ ou d'information de Akaike AIC, doivent maintenant être optimisés dans une procédure de choix de modèle / sélection de variables.

Nous présentons les deux plus simples, la procédure pas à pas ascendante et la procédure pas à pas descendante.

Procédure pas à pas ascendante

Procédure pas à pas ascendante (= **forward stepwise**) :

- On choisit un critère
- On part du modèle nul sans variable
- On effectue les p RLS et on sélectionne le modèle qui optimise le critère choisi
c-à-d qui maximise le $R^2_{\text{ajusté}}$ (resp. minimise le AIC) si tel est le critère choisi
↪ on sélectionne ainsi une première variable \mathcal{V}_{i_1}
- On effectue les $p - 1$ RLM avec \mathcal{V}_{i_1} et une autre variable explicative choisie parmi les $p - 1$ restantes et on sélectionne le modèle qui optimise le critère choisi
↪ on sélectionne ainsi une deuxième variable \mathcal{V}_{i_2}
- On recommence jusqu'à ce qu'on ne puisse plus optimiser le critère
c-à-d que le $R^2_{\text{ajusté}}$ ne puisse plus augmenter (resp. le AIC ne puisse plus diminuer) si tel est le critère choisi

Procédure pas à pas descendante

Procédure pas à pas descendante (= **backward stepwise**) :

- On choisit un critère
- On part du modèle complet à p variables
- On effectue les p RLM à $p - 1$ variables et on sélectionne le modèle qui optimise le critère choisi c-à-d qui maximise le $R_{\text{ajusté}}^2$ (resp. minimise le AIC) si tel est le critère choisi
→ on supprime ainsi une première variable \mathcal{V}_{j_1}
- On effectue les $p - 1$ RLM à $p - 2$ variables choisies parmi celles différentes de \mathcal{V}_{j_1} et on sélectionne le modèle qui optimise le critère choisi
→ on supprime ainsi une deuxième variable \mathcal{V}_{j_2}
- On recommence jusqu'à ce qu'on ne puisse plus optimiser le critère c-à-d que le $R_{\text{ajusté}}^2$ ne puisse plus augmenter (resp. le AIC ne puisse plus diminuer) si tel est le critère choisi

D)4. Robustesse

Pour $i \in \llbracket 1, n \rrbracket$, on dit que le point (x_i, y_i) est un **point influent** s'il a un poids disproportionné, qu'il influe beaucoup sur les paramètres estimés (quand on l'enlève, la droite de régression estimée change beaucoup)

→ il est caractérisé par une distance de Cook $\frac{1}{2} \frac{h_{ii}}{1-h_{ii}} \hat{\epsilon}_i^{(t)}$ élevée (> 1)

Un tel point influent (x_i, y_i) est

- soit un **point levier** → en général, point dans la continuité des autres mais "loin" des autres s'il vérifie un des critères suivants :
 - critères de Welsch : $h_{ii} > \frac{2p}{n}$ ou $(h_{ii} > \frac{3p}{n} \text{ pr } p > 6 \text{ et } n - p > 12)$
 - critère de Huber : $h_{ii} > 0.5$
- soit un **outlier = point aberrant/atypique** → en général, point qui n'est pas "dans la continuité des autres", et est éloigné des régions de confiance/prédiction
si $|\hat{\epsilon}_i^{(t \text{ ext})}| > q_{1-\frac{1}{n}}^{t(n-(p+2))}$ → `outlier.test`
`statsmodels.stats.outliersinfluence.OLSInfluence(reg)`
- soit les deux à la fois !

Annexe : Commande OLS de statsmodels

- `surfaceaugm = statsmodels.api.add_constant(surface) ~` ajout de la colonne de 1
`modelelin = statsmodels.api.OLS(prix,surfaceaugm)`
`reg = modelelin.fit()` ~ initialise le modèle
- `print(reg.summary()) ~`

OLS Regression Results

| | | | |
|-------------------|------------------|---------------------|----------|
| Dep. Variable: | y | R-squared: | 0.842 |
| Model: | OLS | Adj. R-squared: | 0.834 |
| Method: | Least Squares | F-statistic: | 96.26 |
| Date: | Fri, 10 Mar 2023 | Prob (F-statistic): | 1.20e-08 |
| Time: | 18:28:07 | Log-Likelihood: | -103.47 |
| No. Observations: | 20 | AIC: | 210.9 |
| Df Residuals: | 18 | BIC: | 212.9 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|-------|---------|---------|-------|-------|---------|--------|
| const | 33.6438 | 24.445 | 1.376 | 0.186 | -17.713 | 85.001 |
| x1 | 3.8478 | 0.392 | 9.811 | 0.000 | 3.024 | 4.672 |

| | | | |
|----------------|--------|-------------------|-------|
| Omnibus: | 0.506 | Durbin-Watson: | 0.990 |
| Prob(Omnibus): | 0.777 | Jarque-Bera (JB): | 0.583 |
| Skew: | -0.125 | Prob(JB): | 0.747 |
| Kurtosis: | 2.202 | Cond. No. | 151 |

- `reg.params` \leadsto contient le vecteur des paramètres estimés $\hat{\beta}$.
`reg.conf_int(alpha)[j]` \leadsto contient l'intervalle de confiance de niveau $1 - \alpha$ de β_j
`reg.tvalues[j]` \leadsto contient la valeur de la statistique du test de Student de significativité de β_j
`reg.pvalues[j]` \leadsto contient la p-valeur du test de Student de significativité de β_j
- `reg.fittedvalues` \leadsto contient le vecteur des prédictions \hat{y} . (la valeur de \hat{y}_i est à l'indice $i - 1$)
- `reg.resid` \leadsto contient le vecteur des résidus $\hat{\epsilon}$. (la valeur de $\hat{\epsilon}_i$ est à l'indice $i - 1$)
`reg.ssr` \leadsto contient la valeur de la somme des carrés des résidus scr
`reg.ess` \leadsto contient la valeur de la somme des carrés expliquée sce
`reg.centered_tss` \leadsto contient la valeur de la somme des carrés totale sct
`reg.rsquared` \leadsto contient la valeur du coefficient de détermination r^2
- `reg.get_prediction(x_new).conf_int(obs=True, alpha=0.05)`
 \leadsto contient l'intervalle de prévision de Y_{n+1} pour $\mathbf{x_new} = (1, x_{n+1,1}, \dots, x_{n+1,p})$ une nvelle donnée non étiquetée
- `reg.get_prediction(x_new).conf_int(obs=False, alpha=0.05)`
 \leadsto contient l'intervalle de confiance de $\mathbb{E}[Y_{n+1}]$ pour $\mathbf{x_new} = (1, x_{n+1,1}, \dots, x_{n+1,p})$ une nvelle donnée non étiquetée
- `reg.fvalue` \leadsto contient la valeur de la statistique du test de Fisher de significativité du modèle
`reg.f_pvalue` \leadsto contient la p-valeur du test de Fisher de significativité du modèle