HLMA408: Traitement des données

Modèle Linéaire

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Sommaire

Motivation: modèle linéaire et moindres carrés

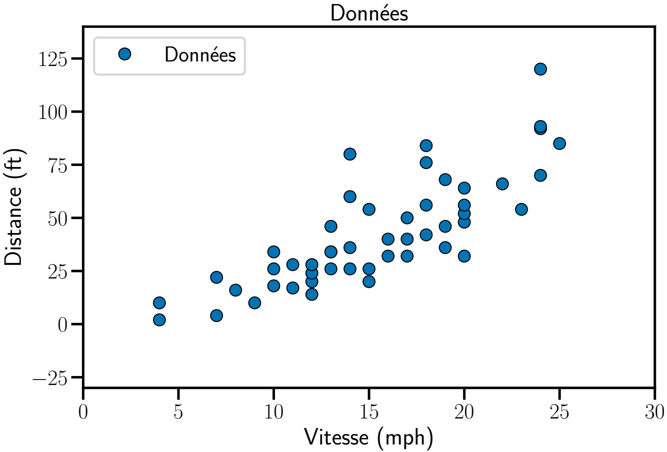
Données : deux variables mesurées / expérience

Étudier :

1. si les variables sont (linéairement) liées
2. quelle est la force du lien
3. si la variable d'intérêt peut être prédite en observant uniquement l'autre

Point de départ en dimension deux

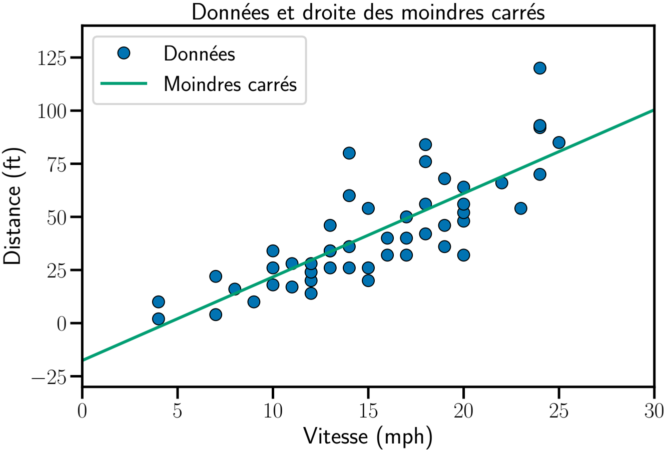
Exemple : vitesse et distance de freinage de voitures; $n = 50$ mesures, vitesse *miles per hour* (mph), distance: feet (ft)



Dataset *cars* : https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/cars.html

Point de départ en dimension deux

Exemple : vitesse et distance de freinage de voitures; $n = 50$ mesures, vitesse *miles per hour* (mph), distance: feet (ft)



Dataset *cars* : https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/cars.html

Sommaire

Sommaire

Modélisation I

Observations: (y_i, x_i), pour $i = 1, \ldots, n$

Hypothèse de modèle linéaire ou de **régression** linéaire:

$$\boxed{y_i \approx \beta_0^\star + \beta_1^\star x_i}$$

- $\beta_0^\star$: ordonnée à l'origine (inconnue)
- $\beta_1^\star$: coefficient directeur (inconnu)

Rem. : les deux paramètres sont inconnus du statisticien

Définition

- y est une **observation** ou une variable à expliquer
- x est une **variable explicative** ou covariable (: *feature*)

Interprétation des notations

Exemple : dataset *cars*

- $n = 50$
- y_i: temps de freinage de la voiture i
- x_i: vitesse de la voiture i
- y: l'observation est le temps de freinage
- x: la variable explicative est la vitesse

L'hypothèse de modèle linéaire :
ici cela revient à postuler que le temps de freinage d'une voiture est **proportionnel** à sa vitesse (!)

Modélisation II

Modèle probabiliste[1] :

$$y_i = \beta_0^\star + \beta_1^\star x_i + \varepsilon_i,$$
$$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \text{ pour } i = 1, \dots, n \ (\sigma^2 \text{ : variance du bruit})$$
$$\beta_0^\star, \beta_1^\star \text{ : paramètres } \textbf{inconnus} \text{ du modèle}$$

Interprétation : $\varepsilon_i = y_i - \beta_0^\star - \beta_1^\star x_i$: erreurs entre le modèle théorique et les observations, représentées par des variables aléatoires ε_i centrées (on parle aussi de **bruit blanc**):

$$\forall i \in [\![1, n]\!], \quad \boxed{\mathbb{E}(\varepsilon_i) = 0}$$

Rem. : l'aspect aléatoire peut avoir diverses causes: bruit de mesure, bruit de transmission, variabilité dans une population, etc.

---

[1] On donne ici un sens au symbole $\approx$ utilisé précédemment

Modélisation III

Modèle probabiliste : $\boxed{y_i = \beta_0^\star + \beta_1^\star x_i + \varepsilon_i}$.

Définition

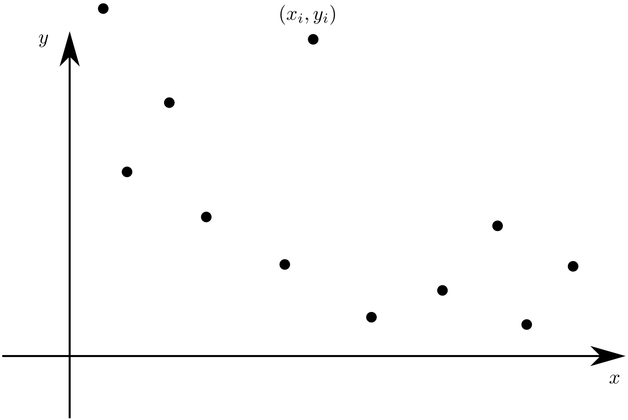
On appelle

- **ordonnée à l'origine** : la quantité $\beta_0^\star$ (: *intercept*)
- **pente** : la quantité $\beta_1^\star$ (: *slope*)

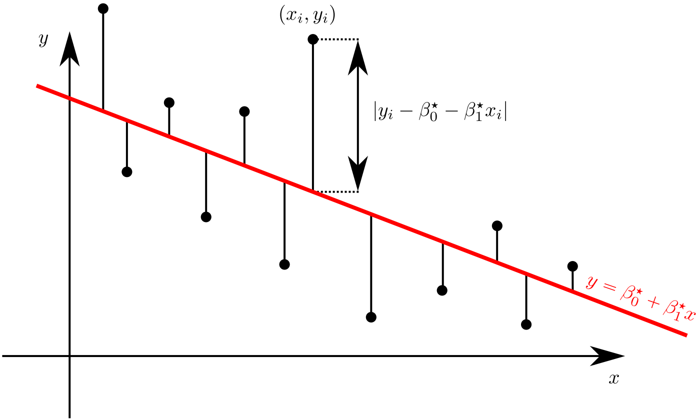
Objectif 1 : Estimer $\beta_0^\star$ et $\beta_1^\star$ (inconnus) par des quantités $\hat{\beta}_0$ et $\hat{\beta}_1$ dépendant des observations (y_i, x_i) pour $i = 1, \ldots, n$.

Objectif 2 : Prédire pour un nouveau point x_{n+1} la valeur non observée y_{n+1} par $\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$

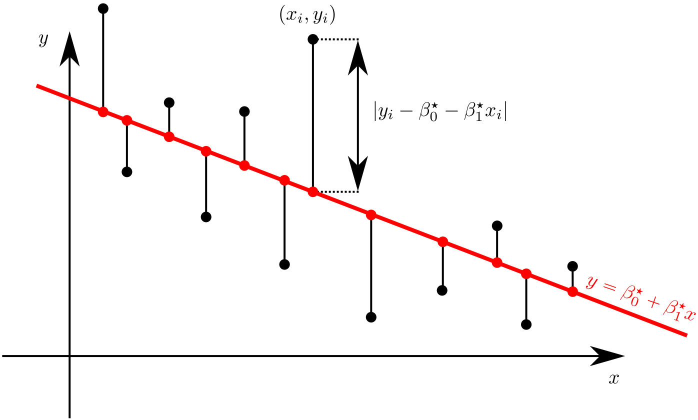
Moindres carrés : visualisation



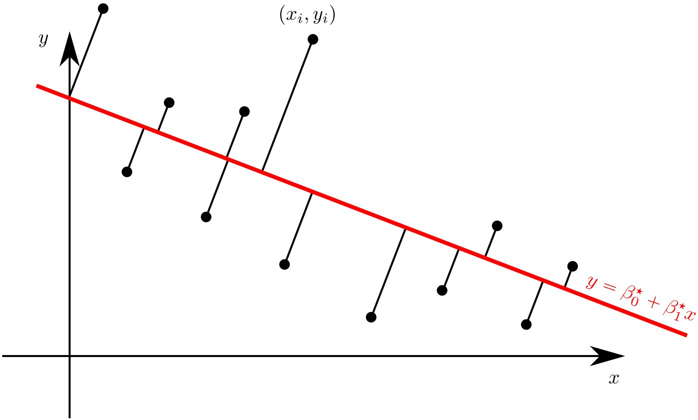
Moindres carrés : visualisation



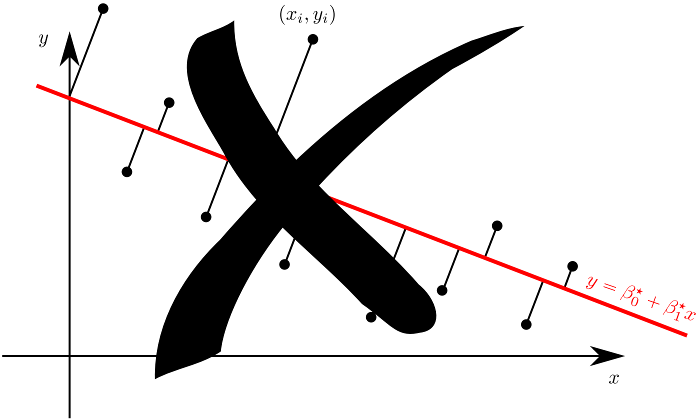
Moindres carrés (totaux) : visualisation



Moindres carrés (totaux) : visualisation



Moindres carrés (totaux) : visualisation



Estimateur des moindres carrés: formulation

Pour des raisons mathématiques (*e.g.*, simplicité computationnelle) on peut choisir de minimiser la somme des carrés des “erreurs”

Définition

L’estimateur des **moindres carrés** est défini comme suit :

$$(\hat{\beta}_0, \hat{\beta}_1) \in \operatorname*{arg\,min}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

- on l’appelle aussi l’estimateur des **moindres carrés ordinaires**, MCO (: *ordinary least-squares, OLS*)
- l’intérêt original vient de ce que les conditions du premier ordre sont équivalentes à résoudre un système linéaire

<u>Rem.</u> : la notation “$\in \operatorname{arg\,min}$” ne présage en rien de l’unicité...

Paternité des moindres carrés

Adrien-Marie Legendre:
"Nouvelles méthodes pour la détermination des orbites des comètes", 1805

Carl Friedrich Gauss:
"Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium" 1809

Aparté

Définition

On définit l'estimateur des **moindres déviations absolues** (🇬🇧 : *Least Absolute Deviation (LAD)*) comme suit:

$$(\hat{\beta}_0, \hat{\beta}_1) \in \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

Rem. : difficile à calculer sans ordinateur; nécessite un algorithme itératif d'optimisation non-lisse (fonctions non différentiables)

Rem. : il est en revanche plus robuste aux points aberrants (🇬🇧 : *outliers*) que l'estimateur des moindres carrés

Paternité des moindres déviations absolues

Ruđer Josip Bošković :"???", 1757

Pierre-Simon de Laplace,
"Traité de mécanique céleste", 1799

Sommaire

Forme explicite des moindres carrés

Théorème

La solution du problème des moindres carrés:

$$\boldsymbol{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \in \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

est unique quand les x_i ne sont pas tous égaux, et s'écrit:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\mathrm{cov}_n(\mathbf{x}, \mathbf{y})}{\mathrm{var}_n(\mathbf{x})} \\ \hat{\beta}_0 = \overline{y}_n - \hat{\beta}_1 \overline{x}_n \end{cases}$$

Rem. : $S_{xx} = n \, \mathrm{var}_n(\mathbf{x})$; $S_{xy} = n \, \mathrm{cov}_n(\mathbf{x}, \mathbf{y})$

Rem : $\mathbf{x} = (x_1, \dots, x_n)^\top$ est non constant

$\iff$ $\mathbf{x}$ n'est pas proportionnel à $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$

$\iff$ $\mathrm{var}_n(\mathbf{x}) \neq 0$

Démonstration :

$$\boxed{\boldsymbol{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \in \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}$$

On cherche donc à minimiser une fonction de deux variables:

$$f(\beta_0, \beta_1) = f(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Conditions nécessaires du premier ordre (CNO):

$$\begin{cases} \frac{\partial f}{\partial \beta_0}(\boldsymbol{\beta}) = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial f}{\partial \beta_1}(\boldsymbol{\beta}) = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

Suite de la démonstration

Rappel : $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et $\overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$

Avec ces notations, les CNO s'écrivent (en divisant par n) :

$$\begin{cases} \frac{\partial f}{\partial \beta_0}(\boldsymbol{\beta}) = -2\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 & \text{(CNO1)} \\ \frac{\partial f}{\partial \beta_1}(\boldsymbol{\beta}) = -2\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 & \text{(CNO2)} \end{cases}$$

$$\iff \begin{cases} \hat{\beta}_0 = \overline{y}_n - \hat{\beta}_1 \overline{x}_n \\ \frac{1}{n}\sum_{i=1}^n x_i y_i = \hat{\beta}_0 \overline{x}_n + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^n x_i^2 \end{cases}$$

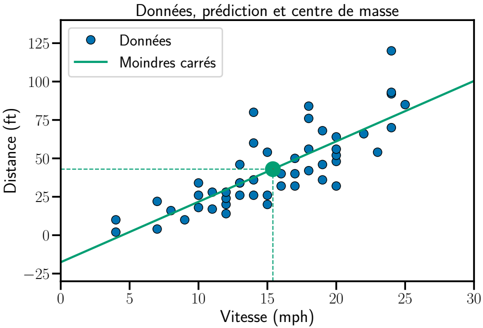
$$\iff \begin{cases} \hat{\beta}_0 = \overline{y}_n - \hat{\beta}_1 \overline{x}_n \\ \frac{1}{n}\sum_{i=1}^n x_i y_i = \overline{y}_n \overline{x}_n - \hat{\beta}_1 (\overline{x}_n)^2 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^n x_i^2 \end{cases}$$

$$\iff \begin{cases} \hat{\beta}_0 = \overline{y}_n - \hat{\beta}_1 \overline{x}_n & \text{(CNO1)} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} = \dfrac{S_{xy}}{S_{xx}} & \text{(CNO2)} \end{cases}$$

Aide : $\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - (\overline{x}_n)^2$

Centre de gravité et interprétation

$$\text{(CNO1)} \iff (\overline{x}_n, \overline{y}_n) \in \{(x,y) \in \mathbb{R}^2 : y = \hat{\beta}_0 + \hat{\beta}_1 x\}$$



- $\overline{vitesse} = 15.4$ mph
- $\overline{distance} = 42.98$ ft
- $\hat{\beta}_0 = -17.58$ ft (l'ordonnée à l'origine)
- $\hat{\beta}_1 = 3.93$ ft/mph (pente de la droite)

Interprétation physique : le **centre de gravité** du nuage de points (c'est le point en vert) est sur la droite de régression (estimée)

Prédiction : la prédiction associée à l'observation moyenne $\overline{x}_n$ est la variable moyenne $\overline{y}_n$

Reformulation vectorielle

Notation : $\mathbf{x} = (x_1, \ldots, x_n)^\top$ et $\mathbf{y} = (y_1, \ldots, y_n)^\top$

$$\text{(CNO2)} \Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$$

$$\text{(CNO2)} \Leftrightarrow \hat{\beta}_1 = \mathrm{corr}_n(\mathbf{x}, \mathbf{y}) \cdot \frac{\sqrt{\mathrm{var}_n(\mathbf{y})}}{\sqrt{\mathrm{var}_n(\mathbf{x})}}$$

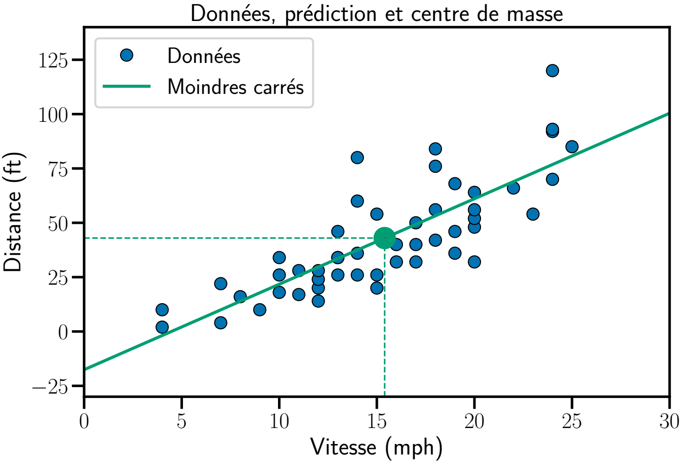
où $$\mathrm{corr}_n(\mathbf{x}, \mathbf{y}) = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\mathrm{var}_n(\mathbf{x})}\sqrt{\mathrm{var}_n(\mathbf{y})}}$$

et $$\mathrm{var}_n(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^n (z_i - \overline{z}_n)^2 \text{ (pour tout } \mathbf{z} = (z_1, \ldots, z_n)^\top\text{)}$$

respectivement **corrélations empiriques** et **variances empiriques**

Retour sur l'exemple du dataset *cars*

Pente de la droite tracée: $\mathrm{corr}_n(\mathbf{x}, \mathbf{y}) \cdot \frac{\sqrt{\mathrm{var}_n(\mathbf{y})}}{\sqrt{\mathrm{var}_n(\mathbf{x})}} = 3.932409.$



Prédictions et prédicteurs

Prédicteur

On appelle **prédicteur** une fonction qui à une nouvelle observation x_{n+1} associe une estimation de la variable à expliquer.
Pour les moindres carrés la prédiction est obtenue par:

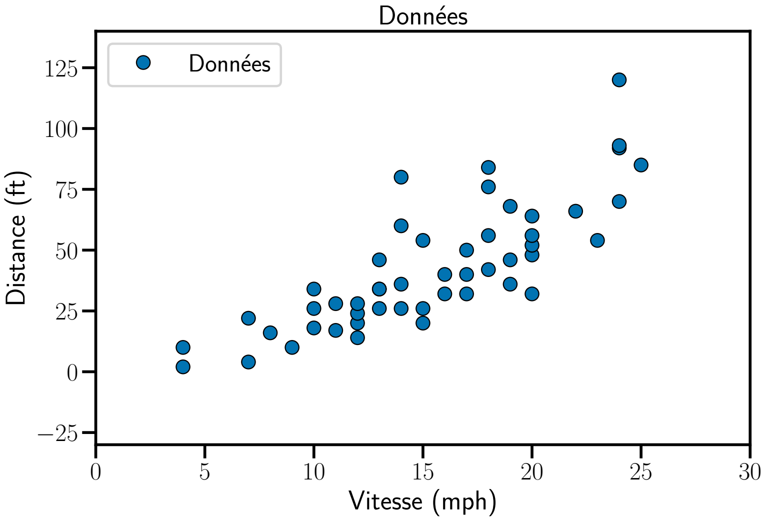
$$\mathrm{pred}(x_{n+1}) = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$

Rem. : souvent on note $\hat{y}_{n+1} = \mathrm{pred}(x_{n+1})$ s'il n'y pas d'ambiguïté

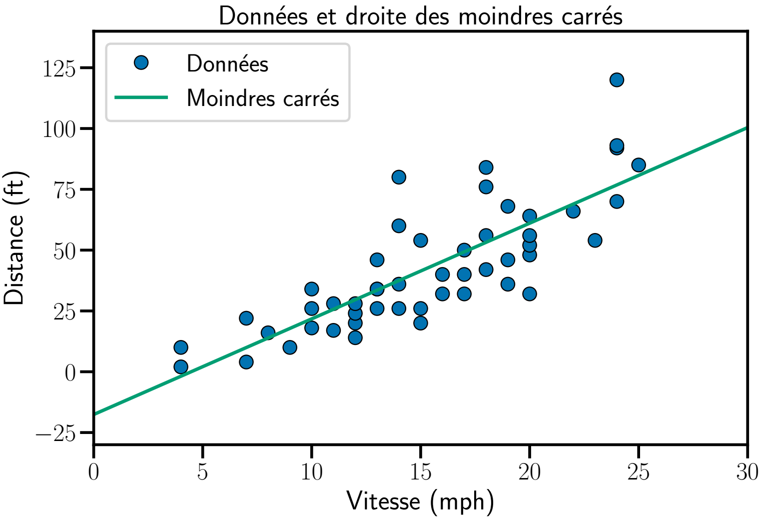
Exemple des voitures : $\hat{\beta}_1 \approx 3.93$ pour $x = 15$ mph, la prédiction est $\hat{y} = -17.58 + 3.93 \times 15 = 41.4$ ft

Notation : on note $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^\top$ le vecteur des prédictions

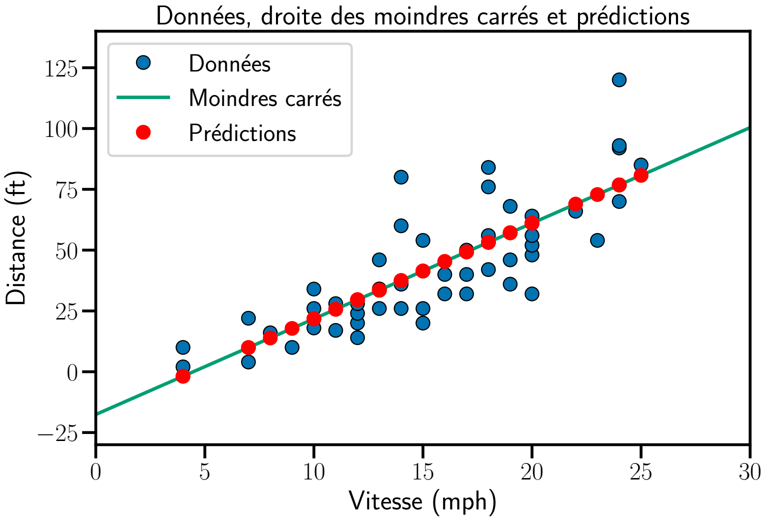
Prédictions: visualisation de $\mathbf{y}$



Prédictions: visualisation de $\mathbf{y}$



Prédictions: visualisation de $\mathbf{y}$



Résidus

Résidus

On appelle **résidu** d'un prédicteur la différence entre la valeur observée et la valeur prédite:

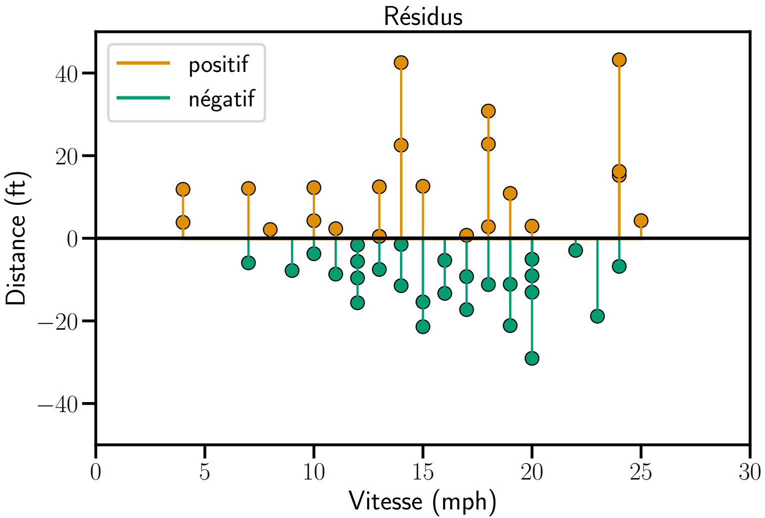
$$\begin{aligned}\hat{\varepsilon}_i =& y_i - \mathrm{pred}(x_i)\\ =& y_i - \hat{y}_i\\ =& y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\end{aligned}$$

Rem. : $\hat{\varepsilon}_i$ peut donc s'interpréter comme un estimateur du bruit ε_i

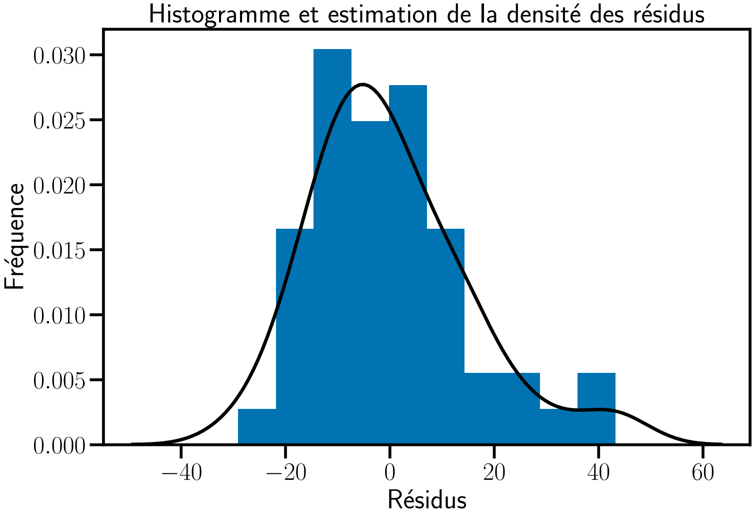
Modèle théorique : $y_i = \beta_0^\star + \beta_1^\star x_i + \varepsilon_i$

Modèle estimé : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$

Résidus


Résidus
positif
négatif
40
20
0
−20
−40
Distance (ft)
0
5
10
15
20
25
30
Vitesse (mph)

Histogramme des résidus



Résidus (suite)

Rappel : $\hat{\varepsilon}_i = y_i - \mathrm{pred}(x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Propriété

Les résidus sont **centrés** (empiriquement): $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Démonstration :

Résidus (suite)

Rappel : $\hat{\varepsilon}_i = y_i - \mathrm{pred}(x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Propriété

Les résidus sont **centrés** (empiriquement): $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Démonstration :

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n} (y_i - \mathrm{pred}(x_i))$$

Résidus (suite)

Rappel : $\hat{\varepsilon}_i = y_i - \text{pred}(x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Propriété

Les résidus sont **centrés** (empiriquement): $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Démonstration :

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \text{pred}(x_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)\end{aligned}$$

Résidus (suite)

Rappel : $\hat{\varepsilon}_i = y_i - \mathrm{pred}(x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Propriété

Les résidus sont **centrés** (empiriquement): $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Démonstration :

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \mathrm{pred}(x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big)
\end{aligned}$$

Résidus (suite)

Rappel : $\hat{\varepsilon}_i = y_i - \mathrm{pred}(x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Propriété

Les résidus sont **centrés** (empiriquement): $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Démonstration :

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \mathrm{pred}(x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big) \\
&= \overline{y}_n - (\hat{\beta}_0 + \hat{\beta}_1 \overline{x}_n) = 0
\end{aligned}
$$

Rem. : la dernière égalité vient de (CNO1)

Deux sources de variabilité des y_i

- variabilité due aux x_i (variable explicative)
- variabilité due aux erreurs ε_i (non-observées), pour un x_i fixé

Définition

Somme des carrés des erreurs (🇬🇧 : *Sum of Squared Errors*) :

$$\mathrm{SSE} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Rem. : on utilise aussi parfois le nom de *Residual Sum of Squares*

Rem. : pour l'analyse mathématique on fait souvent l'hypothèse que les x_i sont déterministes

Estimation de σ^2

Rappel : $y_i = \beta_0^\star + \beta_1^\star x_i + \varepsilon_i$, avec $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2$

Théorème:

L'estimateur de la variance

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2$$

est un estimateur sans biais de σ^2: $\mathbb{E}\left(\hat{\sigma}^2\right) = \sigma^2$.

Rem. : noter que le $n-2$ vient de ce que l'on estime deux paramètres $(\beta_0^\star, \beta_1^\star)$

Exemple des voitures: $\hat{\sigma}^2 \approx 237$ ft^2, *i.e.*, l'estimation de l'écart-type est $\hat{\sigma} \approx 15.4$ ft.

Sommaire

Un point important

Questions:

1. Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont-il proches des valeurs "théoriques" $\beta_0^\star$ et $\beta_1^\star$ du modèle?
2. Quelle est la précision de leur prédiction? Leur biais? Leur variance?

Analyse du biais

Théorème:

Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs non-biaisés de $\beta_0^\star$ et $\beta_1^\star$, c'est-à-dire:

$$\mathbb{E}(\hat{\beta}_0) = \beta_0^\star$$
$$\mathbb{E}(\hat{\beta}_1) = \beta_1^\star$$

Conséquence: le risque quadratique de ces estimateurs est donc leur variance (*cf.* cours biais/variance)

Preuve : calcul du biais

Cas du biais de $\mathbb{E}(\hat{\beta}_1)$:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\left(\frac{\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}\right) \\
&= \frac{\sum_{i=1}^n (x_i - \overline{x}_n)\mathbb{E}\left(y_i - \overline{y}_n\right)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} \qquad (x_i \text{ déterministes})
\end{aligned}$$

Or $\mathbb{E}(y_i) = \beta_0^\star + \beta_1^\star x_i$ et $\mathbb{E}(\overline{y}_n) = \beta_0^\star + \beta_1^\star \overline{x}_n$, donc

$$\mathbb{E}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \overline{x}_n)\beta_1^\star \mathbb{E}\left(x_i - \overline{x}_n\right)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} = \beta_1^\star$$

Cas du biais de $\mathbb{E}(\hat{\beta}_0)$:

$$\mathbb{E}(\hat{\beta}_0) = \mathbb{E}(\overline{y}_n - \hat{\beta}_1 \overline{x}_n) = \beta_0^\star + \beta_1^\star \overline{x}_n - \mathbb{E}(\hat{\beta}_1)\overline{x}_n = \beta_0^\star$$

Variance des estimateurs

Théorème

Les estimateurs

$$\hat{\sigma}_1^2 := \frac{\hat{\sigma}^2}{n} \frac{1}{\mathrm{var}_n(\mathbf{x})} \quad \text{et} \quad \hat{\sigma}_0^2 := \frac{\hat{\sigma}^2}{n} \left(1 + \frac{\overline{x}_n^2}{\mathrm{var}_n(\mathbf{x})} \right)$$

sont des estimateurs sans biais des variances (théoriques) :

$$\mathbb{E}\left[\hat{\sigma}_1^2\right] = \mathbb{V}\mathrm{ar}(\hat{\beta}_1), \qquad \mathbb{E}\left[\hat{\sigma}_0^2\right] = \mathbb{V}\mathrm{ar}(\hat{\beta}_0)$$

où l'on rappelle que $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$

Rem. : le terme $\frac{\hat{\sigma}^2}{n}$ se comporte comme la variance d'une moyenne

Rem. : on retrouve bien les deux sources de variabilité (celle venant des ε_i et celle venant des x_i)

Preuve : voir Cornillon et Matzner-Lober (2011) ou Delyon (2015)

Distribution des estimateurs

- La variable
$$T_1 = \frac{\hat{\beta}_1 - \beta_1^\star}{\hat{\sigma}_1} \sim t(n-2)$$
i.e., suit une loi de Student à $n-2$ degrés de liberté[2]

- La variable
$$T_0 = \frac{\hat{\beta}_0 - \beta_0^\star}{\hat{\sigma}_0} \sim t(n-2)$$
i.e., suit une loi de Student à $n-2$ degrés de liberté

---

[2] de nouveau cela vient du fait que l'on a deux paramètres, $\beta_0^\star$ et $\beta_1^\star$ à estimer dans le modèle

IC sur la pente

Théorème

Un intervalle de confiance de $\beta_1^\star$ au niveau $1-\alpha$ est

$$\left[\hat{\beta}_1 - t_{1-\alpha/2}(n-2)\hat{\sigma}_1, \hat{\beta}_1 + t_{1-\alpha/2}(n-2)\hat{\sigma}_1\right]$$

où $t_{1-\alpha/2}(n-2)$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-2$ degrés de liberté et $\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{n}\frac{1}{\mathrm{var}_n(\mathbf{x})}$

Exemple des voitures: L'intervalle de confiance au niveau 95% est $[3.01, 4.77]$ pour la pente $\beta_1^\star$.

Interprétation : d'après le modèle utilisé, avec une confiance de 95%, en accélérant de 1 mph la voiture, on rallonge en moyenne la distance de freinage d'une distance comprise entre 3.01 et 4.77 ft

IC sur l'ordonnée à l'origine

Théorème

Un intervalle de confiance de $\beta_0^\star$ au niveau $1-\alpha$ est

$$\left[\hat{\beta}_0 - t_{1-\alpha/2}(n-2)\hat{\sigma}_0, \hat{\beta}_0 + t_{1-\alpha/2}(n-2)\hat{\sigma}_0\right]$$

où $t_{1-\alpha/2}(n-2)$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-2$ degrés de liberté et $\hat{\sigma}_0^2 = \frac{\hat{\sigma}^2}{n}\frac{1}{\mathrm{var}_n(\mathbf{x})}$

Exemple des voitures : L'intervalle de confiance au niveau 95% est $[-31.2, -4.99]$ pour la pente $\beta_1^\star$.

Interprétation : le modèle est moins pertinent pour cette quantité, car l'analyse indique qu'à 0mph la voiture met entre -31.2ft et -4.99ft pour s'arrêter.

Prédiction et incertitude

Pour une valeur x fixée et connue on associe

- la valeur (théorique) donnée par le modèle: $\beta_0^\star + \beta_1^\star x$
- la valeur (prédite) donnée par le statisticien: $\hat\beta_0 + \hat\beta_1 x$

Théorème

La variance de $\hat\beta_0 + \hat\beta_1 x$ est estimée sans biais par

$$\frac{\hat\sigma^2}{n}\left(\frac{\mathrm{var}_n(\mathbf{x}) + (x - \overline{x}_n)^2}{\mathrm{var}_n(\mathbf{x})}\right)$$

De plus

$$\frac{(\hat\beta_0 + \hat\beta_1 x) - (\beta_0^\star + \beta_1^\star x)}{\sqrt{\frac{\hat\sigma^2}{n}\left(\frac{\mathrm{var}_n(\mathbf{x}) + (x-\overline{x}_n)^2}{\mathrm{var}_n(\mathbf{x})}\right)}} \sim t(n-2)$$

Sommaire

Objectifs possibles dans le modèle linéaire

Dans le contexte de la régression linéaire, on peut chercher à :

- Tester des hypothèses,
- Construire des intervalles de confiance,
- Faire des prédictions.

Test sur la valeur de $\beta_1^\star$ (pente)

$$\mathcal{H}_0 : \text{“}\beta_1^\star = \mu\text{”} : \text{ valeur connue à tester}$$

Rem. : $\mu = 0$, **cas important** pour tester l'existence d'un lien linéaire entre x et y; si $\mu = 0 \implies y_i = \beta_0^\star + \varepsilon_i$ pour tout $i = 1, \ldots, n$, et donc l'influence(3) (linéaire) de x_i est inexistante

Formalisation du test : $\mathcal{H}_0$: "$\beta_1^\star = \mu$" vs. $\mathcal{H}_1$: "$\beta_1^\star \neq \mu$"

$$T_1 = \frac{\hat{\beta}_1 - \mu}{\hat{\sigma}_1} \sim t(n-2) \quad (\text{sous } \mathcal{H}_0)$$

Rappel : on a $\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{n} \frac{1}{\mathrm{var}_n(\mathbf{x})}$

Conclusion du test :
rejet de $\mathcal{H}_0$ au niveau α si $|T_1| \in \mathcal{R}_\alpha = [t_{1-\alpha/2}(n-2)\ ;\ +\infty[$

(3) Attention : il peut cependant y avoir un lien non linéaire avec x_i, par exemple avec $x_i^2, \exp(x_i)$, etc.

Exemple des voitures

Hypothèse nulle $\mathcal{H}_0$: $\beta_1^\star = 0$.

On a vu $\hat{\beta}_1 = 3.93$ et $\hat{\sigma}^2 = 236.5$.
On estime la variance de $\hat{\beta}_1$ par $\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{n} \frac{1}{\mathrm{var}_n(\mathbf{x})} = 0.17$

Statistique de test : $T_1 = \dfrac{3.93}{\sqrt{0.17}} = 9.53$
Avec $\alpha = 5\%$, $n - 2 = 48$, $t_{0.975}(48) = 9.46$

$\Longrightarrow$ on rejette $\mathcal{H}_0$ car $|T_1| > t_{0.975}(48)$

Intervalle de confiance sur la prédiction en une nouvelle mesure

On observe une nouvelle mesure: $x \in \mathbb{R}$

Théorème

Un IC pour $\beta_0^\star + \beta_1^\star x$ au niveau $1-\alpha$ est

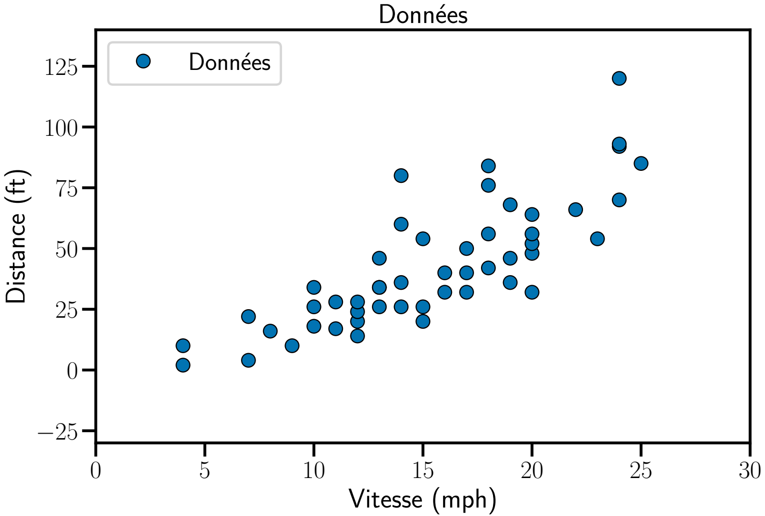
$$\left[\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{1-\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{n}\left(\frac{\mathrm{var}_n(\mathbf{x}) + (x - \overline{x}_n)^2}{\mathrm{var}_n(\mathbf{x})}\right)}\right]$$

où $t_{1-\alpha/2}(n-2)$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-2$ degrés de liberté

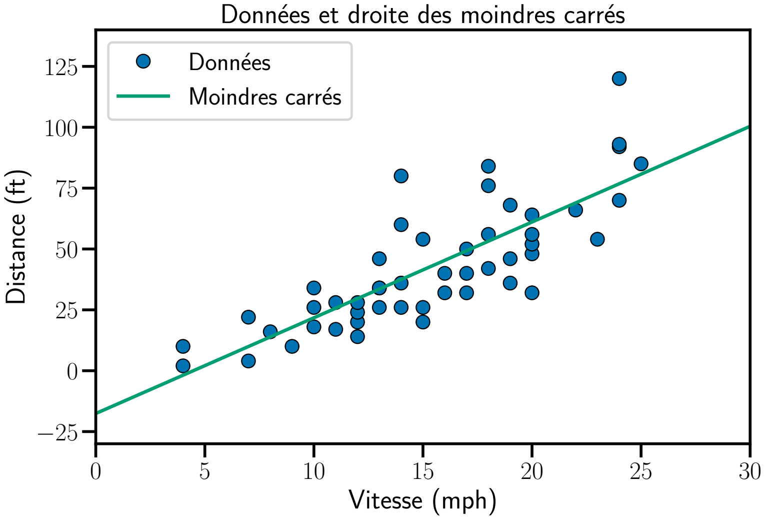
Rem. : l'intervalle de confiance est le plus petit possible quand l'observation x est égale à la moyenne des observations, *i.e.*, $\overline{x}_n$

Rem. : l'intervalle de confiance s'élargit quand on s'éloigne de $\overline{x}_n$

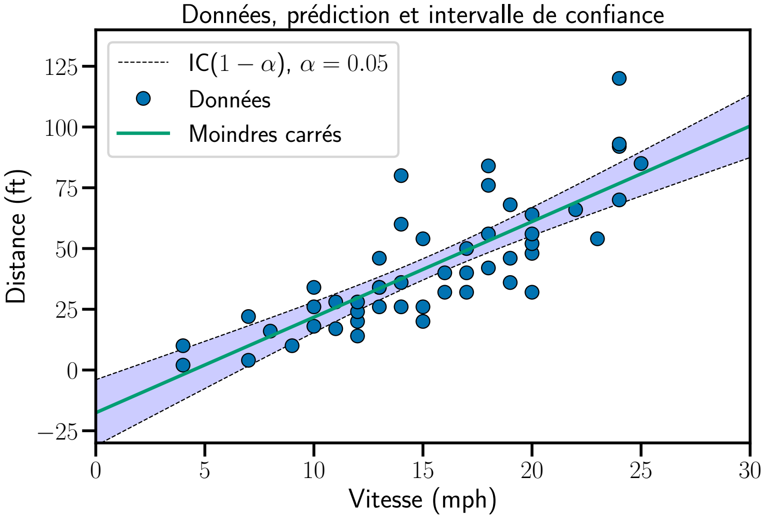
Visualisation de l'intervalle de prédiction



Visualisation de l'intervalle de prédiction



Visualisation de l'intervalle de prédiction



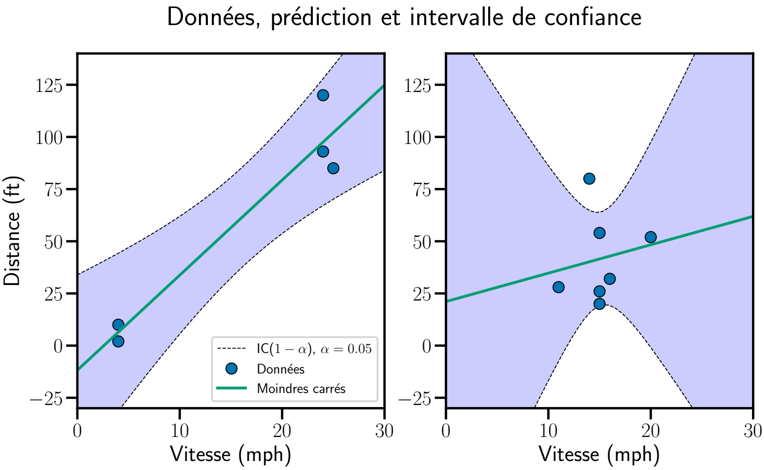
Exemple des voitures

- $\hat{\beta}_1 = 3.93$ ft/mph
- $\hat{\beta}_0 = -17.58$ ft

À la vitesse $x = 15$ mph:
- $\hat{\beta}_0 + \hat{\beta}_1 x = 41.4$ ft
- IC pour la distance d'arrêt : $[37.0, 45.8]$ ft, au niveau 95%.

À la vitesse $x = 5$ mph:
- $\hat{\beta}_0 + \hat{\beta}_1 x = 2.08$ ft
- IC pour la distance d'arrêt : $[-7.6, 11.8]$ au niveau 95%.

Intérêt de la dispersion



Sommaire

Les deux composantes de y

$$\begin{array}{ccccc} \text{Réponse} & = & \text{Valeur expliquée par } x & + & \text{résidu} \\ y_i & = & \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) & + & \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \\ y_i & = & \hat{y}_i & + & \hat{\varepsilon}_i \end{array}$$

Rem. : le vecteur des résidus et le vecteur des prédictions sont orthogonaux, $\sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i = 0$ (conséquences de CNO1 et CNO2).

Théorème

La **somme des carrés des Erreurs**

$$\mathrm{SSE} = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \mathrm{var}_n(\mathbf{y}) - \mathrm{var}_n(\hat{\mathbf{y}}) = \mathrm{var}_n(\mathbf{y}) - \frac{(\mathrm{cov}_n(\mathbf{x},\mathbf{y}))^2}{\mathrm{var}_n(\mathbf{x})}$$

mesure l'écart au modèle linéaire et la **somme des carrés totale**

Décomposition de la variabilité

Théorème

Variabilité de y	$=$	Variabilité expliquée par le modèle	$+$	variabilité résiduelle
$\mathrm{var}_n(\mathbf{y})$	$=$	$\mathrm{var}_n(\hat{\mathbf{y}})$	$+$	$\mathrm{var}_n(\hat{\boldsymbol{\varepsilon}})$
$\mathrm{var}_n(\mathbf{y})$	$=$	$\frac{(\mathrm{cov}_n(\mathbf{x},\mathbf{y}))^2}{\mathrm{var}_n(\mathbf{x})}$	$+$	$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$

Notation :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \hat{\mathbf{y}} = \begin{pmatrix} \hat{y_1} \\ \vdots \\ \hat{y_n} \end{pmatrix} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}, \hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \mathbf{y} - \hat{\mathbf{y}}$$

Preuve du théorème

Comme $\sum_{i=1}^n \hat{y}_i \hat{\varepsilon}_i = 0$, que $\frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}_n$, et que $y_i = \hat{y}_i + \hat{\varepsilon}_i$ pour tout $i \in [\![1, n]\!]$ on obtient

$$\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}_n)^2 = \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - \overline{y}_n)^2 + \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2$$

$$\mathrm{var}_n(\mathbf{y}) = \mathrm{var}_n(\hat{\mathbf{y}}) + \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2$$

Enfin,

$$\begin{aligned}
\mathrm{var}_n(\hat{\mathbf{y}}) &= \mathrm{var}_n(\hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}) \\
&= \mathrm{var}_n(\hat{\beta}_1 \mathbf{x}) \\
&= (\hat{\beta}_1)^2 \,\mathrm{var}_n(\mathbf{x}) \\
&= \left(\frac{\mathrm{cov}_n(\mathbf{x}, \mathbf{y})}{\mathrm{var}_n(\mathbf{x})}\right)^2 \mathrm{var}_n(\mathbf{x}) \\
&= \frac{(\mathrm{cov}_n(\mathbf{x}, \mathbf{y}))^2}{\mathrm{var}_n(\mathbf{x})}
\end{aligned}$$

Coefficient de corrélation linéaire

La **proportion de la variabilité** des y_i **expliquée** par le modèle est

$$r^2 = (\mathrm{corr}_n(\hat{\mathbf{y}}, \mathbf{y}))^2 = (\mathrm{corr}_n(\mathbf{x}, \mathbf{y}))^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Ce nombre est souvent appelé **coefficient R^2 ou r^2**, c'est le carré du coefficient de corrélation linéaire de Spearman

- $r^2 \in [0, 1]$ car $\mathrm{corr}_n(\mathbf{x}, \mathbf{y}) \in [-1, 1]$
- $r^2 = 1 \iff \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 0 \iff$ les points (x_i, y_i) sont sur la droite de régression

Preuve de $(\mathrm{corr}_n(\hat{\mathbf{y}}, \mathbf{y}))^2 = (\mathrm{corr}_n(\mathbf{x}, \mathbf{y}))^2$

$$(\mathrm{corr}_n(\hat{\mathbf{y}}, \mathbf{y}))^2 = \frac{(\mathrm{cov}_n(\hat{\mathbf{y}}, \mathbf{y}))^2}{\mathrm{var}_n(\mathbf{y})\,\mathrm{var}_n(\hat{\mathbf{y}})}$$

Mais avec les relations $\mathrm{var}_n(\hat{\mathbf{y}}) = \frac{(\mathrm{cov}_n(\mathbf{x},\mathbf{y}))^2}{\mathrm{var}_n(\mathbf{x})}$ et $\mathrm{cov}_n(\hat{\mathbf{y}}, \mathbf{y}) = \mathrm{cov}_n(\hat{\beta}_0 + \hat{\beta}_1\mathbf{x}, \mathbf{y}) = \hat{\beta}_1\,\mathrm{cov}_n(\mathbf{x}, \mathbf{y})$

$$\begin{aligned}(\mathrm{corr}_n(\hat{\mathbf{y}}, \mathbf{y}))^2 &= \frac{\hat{\beta}_1\,(\mathrm{cov}_n(\mathbf{x}, \mathbf{y}))^2}{\mathrm{var}_n(\mathbf{y})\frac{(\mathrm{cov}_n(\mathbf{x},\mathbf{y}))^2}{\mathrm{var}_n(\mathbf{x})}} \\ &= \hat{\beta}_1\frac{\mathrm{var}_n(\mathbf{x})}{\mathrm{var}_n(\mathbf{y})}\end{aligned}$$

Et comme (*cf.* slide 21) $\hat{\beta}_1 = \mathrm{corr}_n(\mathbf{x}, \mathbf{y}) \cdot \frac{\sqrt{\mathrm{var}_n(\mathbf{y})}}{\sqrt{\mathrm{var}_n(\mathbf{x})}}$, on obtient:

$$\boxed{(\mathrm{corr}_n(\hat{\mathbf{y}}, \mathbf{y}))^2 = (\mathrm{corr}_n(\mathbf{x}, \mathbf{y}))^2}$$

Quelques remarques sur le modèle de régression linéaire

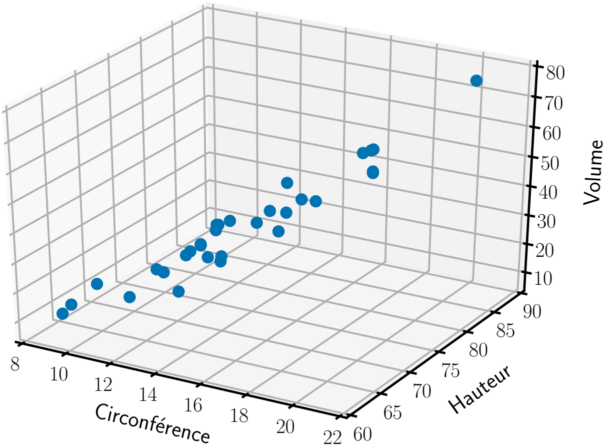
Les hypothèses importantes sont :

- la relation sous-jacente est linéaire
- les bruit/les erreurs ε_i sont indépendantes
- la variance du bruit/de l'erreur est constante pour toutes les observations (modèle homoscédastique[(4)])
- les bruit/les erreurs suivent une loi normale

[(4)] https://fr.wikipedia.org/wiki/Hétéroscédasticité

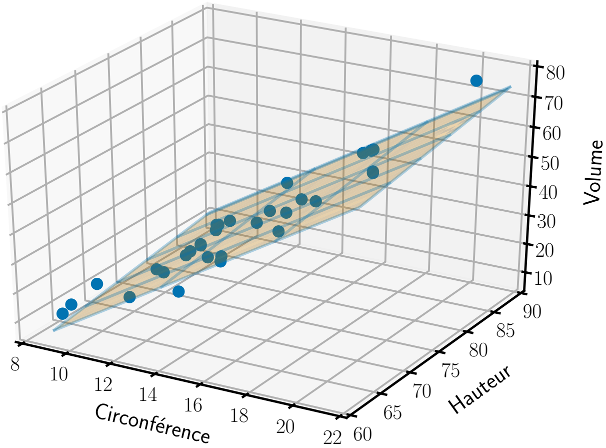
Pour aller plus loin : extension vers un modèle à plusieurs variables explicatives

Exemple : on mesure la taille et la circonférence d'un arbre et on cherche à prédire son volume



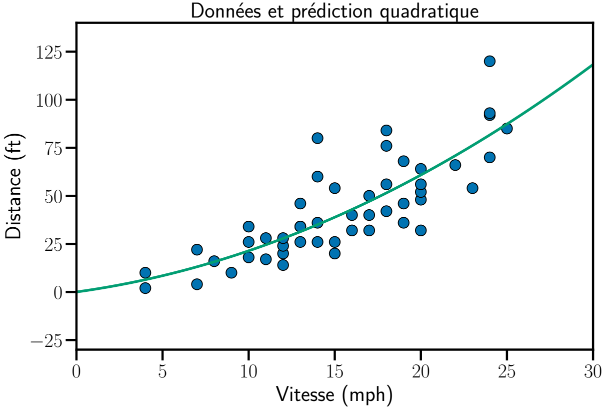
Pour aller plus loin : extension vers un modèle à plusieurs variables explicatives

Exemple : on mesure la taille et la circonférence d'un arbre et on cherche à prédire son volume



Pour aller plus loin : transformation des variables explicatives

Les lois physiques (ou vos souvenirs d'auto-école) conduisent plutôt à choisir une parabole au lieu d'une droite: la même procédure permet d' obtenir l'ajustement suivant en choisissant comme variable explicative x_i^2 au lieu de x_i:



Bibliographie I

- Cornillon, P-A. and E. Matzner-Løber. *Régression avec R*. Springer, Collection Pratique R, 2011, p. 242.
- Delyon, B. "Régression". 2015. URL: https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf.