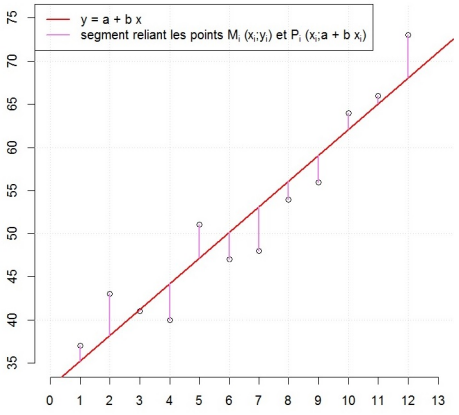
∼ Mémo Régression ∼

Intro : Ajustement d'un nuage de points

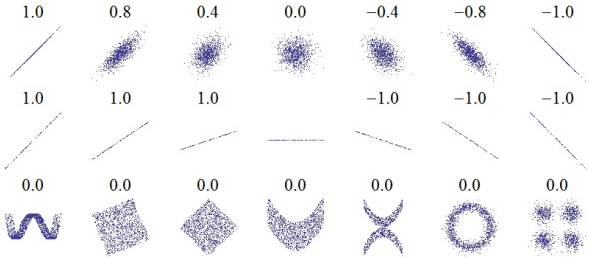
Objectif : On souhaite prédire et expliquer les valeurs d'une variable quantitative Y à partir des valeurs d'une variable quantitative X.
Données : On a n observations de (X, Y) notées $(x_1, y_1), \dots, (x_n, y_n)$.
Estimation : À partir des données, on veut estimer la liaison existante entre Y et X.
Nuage de points : Ensemble des points $\{M_1, \dots, M_n\}$, où M_i est le point de coordonnées (x_i, y_i) dans $\mathbb{R}^2$.
Ajustement affine : Si la silhouette du nuage de points est étirée dans une direction, une relation affine/linéaire entre Y et X est envisageable.
Modèle de régression linéaire ; forme générique : $Y = \alpha + \beta X + \epsilon$, où α et β sont des coefficients inconnus et ϵ est un terme d'erreur.
Estimation : Pour toute valeur x de X, une valeur estimée y de Y est $y = a + bx$, où a et b sont des valeurs estimées de α et β.
Ajustement : À partir des valeurs x de X, estimer avec précision les valeurs de Y correspondantes revient à déterminer a et b de sorte à ce que la droite d'équation $y = a+bx$ ajuste au mieux le nuage de points.
Méthodes d'ajustement : Méthode des points observés, Méthode des points moyens, Méthode des moindres carrés (et plein d'autres).
Méthode des points observés : On considère la droite passant par 2 points, $M_j(x_j, y_j)$ et $M_k(x_k, y_k)$, choisis parmi $M_1, \dots, M_n$. Cette droite est d'équation $y = a + bx$ avec $b = \dfrac{y_k - y_j}{x_k - x_j}$ et $a = y_j - bx_j$.
Méthodes des points moyens : Soient deux ensembles de points du nuage, l'un formé des points les plus à gauche, et l'autre formé des points les plus à droite. On considère la droite passant par les deux points moyens de ces ensembles : $G_1(\overline{x}_1, \overline{y}_1)$ et $G_2(\overline{x}_2, \overline{y}_2)$. Cette droite est d'équation $y = a + bx$, avec $b = \dfrac{\overline{y}_2 - \overline{y}_1}{\overline{x}_2 - \overline{x}_1}$ et $a = \overline{y}_1 - b\overline{x}_1$.

Méthodes des moindres carrés (ordinaires) : On considère la droite d'équation $y = a + bx$, avec a et b qui rendent minimale la somme des carrés : $\displaystyle\sum_{i=1}^{n}(y_i - (a + bx_i))^2$.
Idée : Minimiser la somme des carrés des distances $d_i = |y_i - (a+bx_i)|$, longueur entre $M_i(x_i, y_i)$ et le point $P_i(x_i, a + bx_i)$:



Outils : $\text{sce}_x = \displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2 = (n-1)s_x^2$, $\text{sce}_y = \displaystyle\sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1)s_y^2$, $\text{spe}_{x,y} = \displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}$.
Formules : On a $b = \dfrac{\text{spe}_{x,y}}{\text{sce}_x}$ et $a = \overline{y} - b\overline{x}$. La droite d'équation $y = a + bx$ est appelée droite de régression.
Coefficient de corrélation linéaire : $r_{x,y} = \dfrac{\text{spe}_{x,y}}{\sqrt{\text{sce}_x \text{sce}_y}}$, $r_{x,y} \in [-1, 1]$.
Plus $|r_{x,y}|$ est proche de 1, plus la liaison linéaire entre Y et X est forte.
On a $b = \dfrac{s_y}{s_x} r_{x,y}$ et $|r_{x,y}| = \sqrt{1 - \dfrac{1}{\text{sce}_y}\displaystyle\sum_{i=1}^{n}(y_i - (a + bx_i))^2}$.

Exemples de nuages de points et des valeurs de $r_{x,y}$ associées :



Ajustement non-affine : Un nuage de points peut être visuellement mieux ajusté par une courbe que par une droite.
Modèle de régression non-linéaire ; forme générique :
$g(Y) = \alpha + \beta f(X) + \epsilon$, où α et β sont des coefficients inconnus, f et g sont des fonctions à choisir et ϵ est un terme d'erreur.
Choix des fonctions : On choisit f et g de sorte à ce que le nuage de points $\{(f(x_1), g(y_1)), \dots, (f(x_n), g(y_n))\}$ ait une silhouette très étirée dans une direction.
Estimation : Pour toute valeur x de X, une valeur estimée y de Y vérifie $g(y) = a + bf(x) \Leftrightarrow y = g^{-1}(a + bf(x))$, où a et b sont des valeurs estimées de α et β.
Méthodes : Une des méthodes introduites précédemment avec le nouveau nuage de points $\{(f(x_1), g(y_1)), \dots, (f(x_n), g(y_n))\}$.

Transition

Méthode des moindres distances : On considère la droite d'équation $y = a+bx$, avec a et b qui rendent minimale la somme des carrés $\displaystyle\sum_{i=1}^{n} d_i^2$, avec d_i la distance entre le point $M_i(x_i, y_i)$ et le point sur la droite mesurée de façon perpendiculaire. Il y a bien d'autres méthodes ...
Question : Pourquoi se focaliser sur la méthode des moindres carrés ?

- Les formules de a et b sont aisément calculables.
- On a des garanties théoriques sur la précision de l'ajustement.
- Elle s'étend à plusieurs "variables explicatives" $X_1, \dots, X_p$.

Estimateur des moindres carrés ordinaires (*emco*)

Objectif : On souhaite prédire et expliquer les valeurs d'une variable quantitative Y à partir des valeurs de p variables $X_1, \dots, X_p$.
Vocabulaire : On veut "expliquer Y à partir de $X_1, \dots, X_p$", Y est une "variable à expliquer" et $X_1, \dots, X_p$ sont des "variables explicatives".
Données : On a n observations de $(Y, X_1, \dots, X_p)$ notées $(y_1, x_{1,1}, \dots, x_{p,1}), \dots, (y_n, x_{1,n}, \dots, x_{p,n})$:

Y	X_1	$\dots$	X_p
y_1	$x_{1,1}$	$\dots$	$x_{p,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{1,n}$	$\dots$	$x_{p,n}$

Estimation : À partir des données, on veut estimer la liaison existante entre Y et $X_1, \dots, X_p$.
Modèle de régression linéaire multiple (*rlm*) ; forme générique :
$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \epsilon$, où $\beta_0, \dots, \beta_p$ sont des coefficients inconnus et ϵ est un terme d'erreur.
Objectif : Estimer convenablement $\beta_0, \dots, \beta_p$ à partir des données.
Modèle de *rlm* : On modélise les variables comme des *var*,

- $(x_{1,i}, \dots, x_{p,i})$ est une réalisation de $(X_1, \dots, X_p)$,
- sachant que $(X_1, \dots, X_p) = (x_{1,i}, \dots, x_{p,i})$, y_i est une réalisation de $Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \epsilon_i$, où ϵ_i est une *var* indépendante de $X_1, \dots, X_p$ avec $\mathbb{E}(\epsilon_i) = 0$.

Écriture matricielle : $Y = X\beta + \epsilon$,
$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$, $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$.
Emco de β : On pose $\|x\|^2 = x^t x$. Sous l'hypothèse qu'il soit unique, l'*emco* de β est le $\widehat{\beta}$ qui rend minimale la somme des carrés : $\|Y - X\widehat{\beta}\|^2 \Leftrightarrow \widehat{\beta} = (X^t X)^{-1} X^t Y$.
Emco de β_j : $\widehat{\beta}_j = [\widehat{\beta}]_{j+1}$ est l'*emco* de β_j.

Valeur moyenne : La valeur moyenne de Y lorsque $(X_1,\dots,X_p) = (x_1,\dots,x_p) = x$ est : $y_x = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.
Estimateur de y_x : Un estimateur de y_x est $\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \dots + \widehat{\beta}_p x_p$.

Emco ponctuel de β : $b = (X^tX)^{-1}X^ty$ avec $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.

Emco ponctuel de β_j : $b_j = [b]_{j+1}$ est l'*emco* ponctuel de β.
Estimation ponctuelle de y_x : $d_x = b_0 + b_1x_1 + \dots + b_px_p$. On dit que d_x est la valeur prédite de Y quand $(X_1,\dots,X_p) = x$.

Coefficient de détermination : $R^2 = 1 - \dfrac{\|Xb - y\|^2}{\|\overline{y}1_n - y\|^2}$. On a $R^2 \in [0,1]$.

Plus R^2 est proche de 1, plus la liaison linéaire entre Y et $X_1,\dots,X_p$ est forte, plus le modèle de *rlm* est pertinent.

Coefficient de détermination ajusté : $\overline{R}^2 = 1 - \dfrac{n-1}{n-(p+1)}(1-R^2)$.

Retour sur le modèle de régression linéaire simple (*rls*)

Modèle de *rls* : Modèle de *rlm* avec $p = 1$: $Y = \beta_0 + \beta_1 X_1 + \epsilon$.
Emco ponctuel de β_j : $b = (X^tX)^{-1}X^ty \Rightarrow b_1 = \dfrac{\text{spe}_{x,y}}{\text{sce}_x}$, $b_0 = \overline{y} - b_1\overline{x}_1$.
Estimation de y_x : $d_x = b_0 + b_1x_1$.
Droite de régression : Droite d'équation : $y = b_0 + b_1x$.
Coefficient de corrélation linéaire : $r_{x,y} = \dfrac{\text{spe}_{x,y}}{\sqrt{\text{sce}_x\text{sce}_y}}$.

On a $r_{x,y} \in [-1,1]$, $b_1 = \dfrac{s_y}{s_x}r_{x,y}$ et $r^2_{x,y} = R^2 \Rightarrow$ même rôle que R^2.

Propriétés standards et lois associées

Hypothèses standards : On suppose que

- X est de rang colonnes plein,
- ϵ et $X_1,\dots,X_p$ sont indépendantes,
- $\epsilon \sim \mathcal{N}_n(0_n, \sigma^2\mathbb{I}_n)$ où $\sigma > 0$ est un paramètre inconnu.

Loi de Y : $Y \sim \mathcal{N}_n\left(X\beta, \sigma^2\mathbb{I}_n\right)$.
Loi de $\widehat{\beta}$: $\widehat{\beta} \sim \mathcal{N}_{p+1}\left(\beta, \sigma^2(X^tX)^{-1}\right)$. En particulier :

- $\widehat{\beta}$ est un estimateur sans biais de β : $\mathbb{E}_{p+1}(\widehat{\beta}) = \beta$,
- $\mathbb{V}_{p+1}(\widehat{\beta}) = \sigma^2(X^tX)^{-1}$,
- $\widehat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2[(X^tX)^{-1}]_{j+1,j+1}\right)$,
- si $[(X^tX)^{-1}]_{j+1,k+1} = 0$, alors $\widehat{\beta}_j$ et $\widehat{\beta}_k$ sont indépendantes.

Emv et *Emco* : L'*emco* $\widehat{\beta}$ est l'*emv* de β. Il est donc fortement consistant et asymptotiquement efficace.
Estimateur de σ^2 : Un estimateur sans biais de σ^2 est $\widehat{\sigma}^2 = \dfrac{1}{n-(p+1)}\|Y - X\widehat{\beta}\|^2$. Il vérifie :

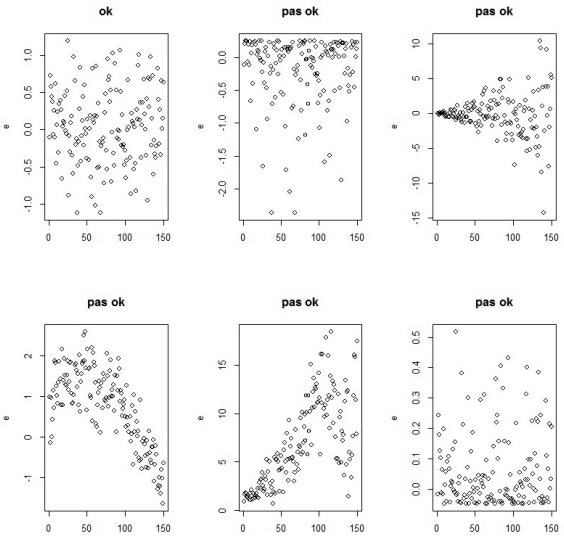
- $\widehat{\sigma}^2$ et $\widehat{\beta}$ sont indépendantes,
- $(n-(p+1))\dfrac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$.

Degré de liberté ν : On pose $\nu = n - (p+1)$.
Emco et loi de Student :
$\dfrac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}} \sim \mathcal{T}(\nu)$, $\dfrac{\widehat{Y}_x - y_x}{\widehat{\sigma}\sqrt{x_\bullet(X^tX)^{-1}x_\bullet^t}} \sim \mathcal{T}(\nu)$.
Emco et loi de Fisher : Soit Q une matrice à $p+1$ colonnes et k lignes : $\dfrac{(Q\widehat{\beta} - Q\beta)^t(Q(X^tX)^{-1}Q^t)^{-1}(Q\widehat{\beta} - Q\beta)}{k\widehat{\sigma}^2} \sim \mathcal{F}(k,\nu)$.
Estimations ponctuelles : Une estimation ponctuelle de

- σ est $s = \sqrt{\dfrac{1}{n-(p+1)}\|y - Xb\|^2}$,
- l'écart-type de $\widehat{\beta}_j$ est $\text{ete}_j = s\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}$,
- l'écart-type de $\widehat{Y}_x$ est $\text{ete}_x = s\sqrt{x_\bullet(X^tX)^{-1}x_\bullet^t}$.

Résidus : $e_1,\dots,e_n$ avec $e_i = y_i - d_{x_i}$.
Graphique des résidus : Nuage de points $\mathcal{N}_e = \{(1,e_1),\dots,(n,e_n)\}$.

Validation des hypothèses standards avec le graphique des résidus :



Re(re)tour sur le modèle de *rls*

Lois de $\widehat{\beta}_j$: $\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma^2\dfrac{1}{\text{sce}_x}\right)$, $\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\dfrac{1}{n} + \dfrac{\overline{x}_1^2}{\text{sce}_x}\right)\right)$.
Loi de $\widehat{Y}_x$: $\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1x_1 \sim \mathcal{N}\left(y_x, \sigma^2\left(\dfrac{1}{n} + \dfrac{(x_1 - \overline{x}_1)^2}{\text{sce}_x}\right)\right)$.
Estimations ponctuelles : Une estimation ponctuelle de

- σ est $s = \sqrt{\dfrac{1}{n-2}\|y - Xb\|^2} = \sqrt{\dfrac{(n-1)s_y^2(1-r_{x,y}^2)}{n-2}}$,
- l'écart-type de $\widehat{\beta}_1$ est $\text{ete}_1 = s\sqrt{\dfrac{1}{\text{sce}_x}}$,
- l'écart-type de $\widehat{\beta}_0$ est $\text{ete}_0 = s\sqrt{\dfrac{1}{n} + \dfrac{\overline{x}_1^2}{\text{sce}_x}}$,
- l'écart-type de $\widehat{Y}_x = \widehat{\beta}_0 + \widehat{\beta}_1x_1$ est $\text{ete}_x = s\sqrt{\dfrac{1}{n} + \dfrac{(x_1 - \overline{x}_1)^2}{\text{sce}_x}}$.

Intervalles et volume de confiance (niveau $100(1-\alpha)\%$)

Quantité $t_\alpha(\nu)$: $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, $T \sim \mathcal{T}(\nu)$.
Intervalle de confiance pour β_j : $i_{\beta_j} = [b_j - t_\alpha(\nu)\text{ete}_j,\ b_j + t_\alpha(\nu)\text{ete}_j]$.
Intervalle de confiance pour y_x : $i_{y_x} = [d_x - t_\alpha(\nu)\text{ete}_x,\ d_x + t_\alpha(\nu)\text{ete}_x]$.
Volume de confiance pour $Q\beta$: Soit Q une matrice à $p+1$ colonnes et k lignes. $v_{Q\beta} = \{u \in \mathbb{R}^{p+1}; (Qb - Qu)^t(Q(X^tX)^{-1}Q^t)^{-1}(Qb - Qu) \leq ks^2f_\alpha(k,\nu)\}$, $\mathbb{P}(F \geq f_\alpha(k,\nu)) = \alpha$, $F \sim \mathcal{F}(k,\nu)$.

Tests statistiques (risque $100\alpha\%$)

Tests statistiques pour β_j :

H_1	p-valeurs
$\beta_j \neq r$	$\mathbb{P}(\lvert T\rvert \geq \lvert t_{obs}\rvert)$
$\beta_j > r$	$\mathbb{P}(T \geq t_{obs})$
$\beta_j < r$	$\mathbb{P}(T \leq t_{obs})$

, $t_{obs} = \dfrac{b_j - r}{\text{ete}_j}$, $T \sim \mathcal{T}(\nu)$. Par exemple, si $r = 0$, $H_1 : \beta_j \neq 0$ et p-valeur $\in]0.001,\ 0.01]$; **, l'influence de X_j sur Y est très significative.
Test de Fisher : Soit Q une matrice à $p+1$ colonnes et k lignes, H_1 : $Q\beta \neq r$, p-valeur $= \mathbb{P}(F \geq f_{obs})$,
$f_{obs} = \dfrac{(Qb - r)^t(Q(X^tX)^{-1}Q^t)^{-1}(Qb - r)}{ks^2}$, $F \sim \mathcal{F}(k,\nu)$.
Test global de Fisher : H_1 : "il y a au moins un coefficient β_j non nul", p-valeur $= \mathbb{P}(F \geq f_{obs})$, $f_{obs} = \dfrac{R^2}{1-R^2}\dfrac{n-(p+1)}{p}$, $F \sim \mathcal{F}(p,\nu)$.
Comparaison de modèles : Λ un sous-ensemble de $\{1,\dots,p\}$ ayant k éléments, H_0 : "$\beta_j = 0$ pour tout $j \in \Lambda$", p-valeur $= \mathbb{P}(F \geq f_{obs})$, $f_{obs} = \dfrac{\|X_\Lambda b_\Lambda - Xb\|^2}{ks^2}$, $b_\Lambda = (X_\Lambda^tX_\Lambda)^{-1}X_\Lambda^ty$, $F \sim \mathcal{F}(k,\nu)$. S'il y a non rejet de H_0, on admet que l'on peut enlever $(X_j)_{j\in\Lambda}$ du modèle.