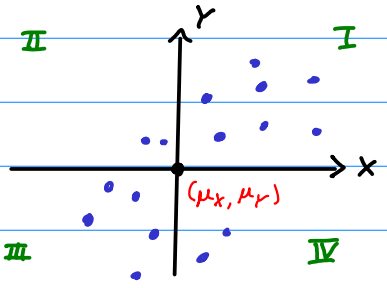
Régression linéaire

But : Établir un lien linéaire entre deux variables aléatoire X et Y.

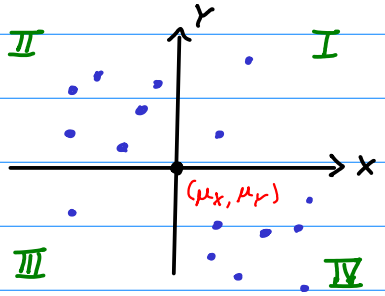
§ Retour sur la covariance.

$$\mathrm{Cov}(X,Y) \overset{\text{déf}}{=} \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

où $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$.



$\mathrm{Cov}(X,Y) > 0$ tendance à se trouver dans les quadrants I et III.



$\mathrm{Cov}(X,Y) < 0$ tendance à se trouver dans les quadrants II et IV.

Estimateur de la covariance : Soit $(X_1,Y_1), \dots, (X_n,Y_n)$ un échantillon iid d'une population de distribution (X,Y). Il est naturel de prendre

$$\widehat{\mathrm{cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}), \quad \text{où} \quad \bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j, \quad \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

Vérifion qu'il est sans biais. D'abord, on a

$$\mathbb{E}\left[(X_i-\bar{X})(Y_i-\bar{Y})\right] = \mathbb{E}\left[\left((X_i-\mu_X)+(\mu_X-\bar{X})\right)\left((Y_i-\mu_Y)+(\mu_Y-\bar{Y})\right)\right]$$

$$= \mathbb{E}\left[(X_i-\mu_X)(Y_i-\mu_Y)\right] - \mathbb{E}\left[(X_i-\mu_X)(\bar{Y}-\mu_Y)\right]$$

$$- \mathbb{E}\left[(\bar{X}-\mu_X)(Y_i-\mu_Y)\right] + \mathbb{E}\left[(\bar{X}-\mu_X)(\bar{Y}-\mu_Y)\right]$$

$$= \mathrm{Cov}(X_i,Y_i) - \underset{(1)}{\mathrm{Cov}(X_i,\bar{Y})} - \underset{(2)}{\mathrm{Cov}(\bar{X},Y_i)} + \underset{(3)}{\mathrm{Cov}(\bar{X},\bar{Y})}$$

$\mathrm{Cov}(X_i, Y_j) = 0$ pour $i \neq j$ car elles sont indépendantes

① $$\mathbb{C}ov\left(X_i, \frac{1}{n}\sum_{j=1}^{n} Y_j\right) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{C}ov(X_i, Y_j) = \frac{1}{n}\left(0 + \cdots + 0 + \mathbb{C}ov(X_i, Y_i) + 0 + \cdots + 0\right)$$

$$= \frac{1}{n}\mathbb{C}ov(X_i, Y_i) = \frac{1}{n}\mathbb{C}ov(X, Y)$$

② Comme le ①

③ $$\mathbb{C}ov(\bar{X}, \bar{Y}) = \mathbb{C}o\left[\frac{1}{n}\sum_{j=1}^{n} X_j, \frac{1}{n}\sum_{m=1}^{n} Y_m\right] = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{m=1}^{n} \mathbb{C}ov(X_j, Y_m)$$

$$= \frac{1}{n^2}\sum_{j=1}^{n} \mathbb{C}ov(X_j, Y_j) = \frac{1}{n}\mathbb{C}ov(X, Y)$$

On a donc

$$\mathbb{E}\left[(X_i - \bar{X})(Y_i - \bar{Y})\right] = \mathbb{C}ov(X, Y) - \frac{1}{n}\mathbb{C}ov(X, Y) - \frac{1}{n}\mathbb{C}ov(X, Y) + \frac{1}{n}\mathbb{C}ov(X, Y)$$

$$= \frac{n-1}{n}\mathbb{C}ov(X, Y)$$

Enfin, on a

$$\mathbb{E}\left[\widehat{cov}(X, Y)\right] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\right]$$

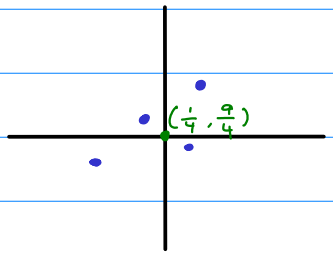
$$= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \bar{X})(Y_i - \bar{Y})\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{n-1}{n}\mathbb{C}ov(X, Y) = \mathbb{C}ov(X, Y).$$

Pour calculer la covariance empirique d'un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$, on fera $\widehat{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Ex. $(1,3)$, $(1,2)$, $(-1,1)$, $(0,3)$

$\bar{x} = \frac{1+1-1+0}{4} = \frac{1}{4}$ $\quad \bar{y} = \frac{3+2+1+3}{4} = \frac{9}{4}$



i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
1	0,75	0,75	0,5625
2	0,75	-0,25	-0,1875
3	-1,25	-1,25	1,5625
4	-0,25	0,75	-0,1875
			1,175 Somme

$\widehat{Cov}(\hat{x}, \hat{y}) = \frac{1,75}{3} = 0,58\overline{3}$

§ Corrélation

$$\rho = Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Rappel : $-1 \leq Cor(X,Y) \leq 1$ et $Cor(X,Y) = \pm 1$ ssi $\exists a, b$, $\mathbb{P}(Y = aX + b) = 1$

$Cor(X,Y)$ est une mesure (sans unité) de la force du lien linéaire entre X et Y.

Un estimateur naturel de la corrélation est

$$\hat{r} = \widehat{Cor}(X,Y) = \frac{\hat{cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

(Les $\frac{1}{n-1}$ se simplifient)

mais celui-ci est biaisé.

Il est difficile de trouver un estimateur simple et sans biais. En pratique, on utilise plutôt

$$\hat{\rho} = r + \frac{r(1-r^2)}{2(n-4)}$$

qui est peu près sans biais lorsque $n \geq 18$ et la distribution de la population est normalement distribuée.

Pour le cours, on se contentera de $\hat{r}$, qui est plus simple à calculer.

<u>Exemple de calcul</u> Calculer la corrélation empirique de l'échantillon

$(1,-1)$ $(0,1)$ $(-\frac{1}{2},3)$ $(1,2)$ $(\frac{1}{2},\frac{1}{2})$

pas besoin de diviser par $\frac{1}{n-1}$

Il faut calculer : $\bar{x}$, $\bar{y}$, $\Sigma(x_i-\bar{x})^2$, $\Sigma(y_i-\bar{y})^2$, $\Sigma(x_i-\bar{x})(y_i-\bar{y})$

$$\bar{x} = \frac{1+0-\frac{1}{2}+1+\frac{1}{2}}{5} = \frac{2}{5} = 0,4 \qquad \bar{y} = \frac{-1+1+3+2+\frac{1}{2}}{5} = \frac{5,5}{5} = 1,1$$

i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	0,6	0,36	-2,1	4,41	-1,26
2	-0,4	0,16	-0,1	0,01	0,04
3	-0,9	0,81	1,9	3,61	-1,71
4	0,6	0,36	0,9	0,81	0,54
5	0,1	0,01	-0,6	0,36	-0,06
Σ		1,7		9,2	-2,45

$$\hat{r} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(\Sigma(x_i-\bar{x})^2)(\Sigma(y_i-\bar{y})^2)}} = \frac{-2,45}{\sqrt{1,7\times 9,2}} = -0,6195$$

§ Loi normale multidimensionnelle

On fait seulement le cas en dimension 2.

On définit la <u>matrice de covariance</u> d'un vecteur aléatoire (X,Y) par

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix}$$

C'est une matrice symétrique. Son déterminant est

$$\begin{aligned}\det \Sigma &= \mathrm{Var}(X)\mathrm{Var}(Y) - \mathrm{Cov}(X,Y)^2 \\ &= \mathrm{Var}(X)\mathrm{Var}(Y)\left[1 - \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)\mathrm{Var}(Y)}\right] \\ &= \mathrm{Var}(X)\mathrm{Var}(Y)\left(1 - \mathrm{Cor}(X,Y)^2\right)\end{aligned}$$

La loi normale multidimentionnelle $N(\vec{\mu}, \Sigma)$ est distribuée selon $\vec{\mu} = \begin{pmatrix}\mu_x \\ \mu_y\end{pmatrix}$ la fonction de densité

$$f(\vec{x}) = \frac{1}{2\pi\sqrt{\det\Sigma}}\, e^{-\frac{1}{2}\vec{x}\,\Sigma^{-1/2}\,\vec{x}^t} \qquad \text{où } \vec{x} = \begin{pmatrix}x \\ y\end{pmatrix}$$

après simplification

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y}\right)}$$

Dans le cas où $\mu_x = 0$ et $\mu_y = 0$, la densité devient

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)$$

La corrélation apparaît de façon naturelle dans cette distribution importante.

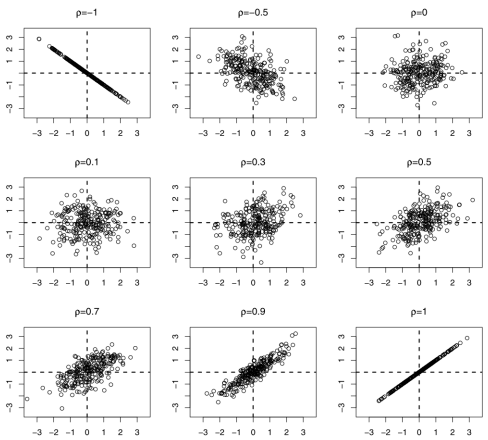
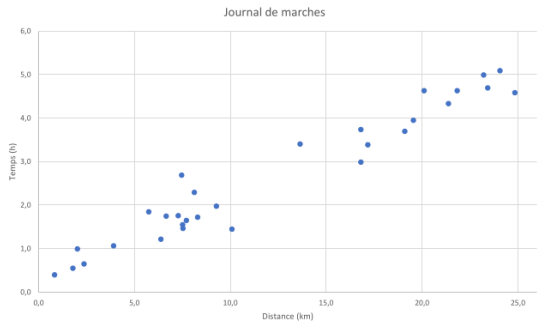


Diagramme de dispersion de $N(\vec{0}, \Sigma)$ avec différentes valeurs de corrélation ρ.

(Source : Rousson, V. Statistique appliquée aux sciences de la vie)

§ Régression linéaire.

Un <u>diagramme de dispersion</u> est un nuage de points dans le plan-XY où chaque point (x,y) est une donnée observée.



(Source : Note de cours d'Élise Darignon)

Soit (X,Y) un vecteur aléatoire. On cherche une relation linéaire (affine) entre X et Y : $Y = aX+b$. On appelle X la <u>variable explicative</u> et Y, la <u>variable de réponse</u>.

On a vu que si $Cor(X,Y) = \pm 1$, alors $Y = aX+b$. En général, cette relation ne sera pas parfaite. On a plutôt

$$Y = aX + b + e \quad , \quad \text{où } e \text{ est un terme d'erreur aléatoire appelé } \underline{\text{résidu}}$$

Le but est de trouver a et b (des constantes) qui minimise l'erreur.

- Cas parfait : si $e=0$, on a $Cov(X,Y) = a\,\mathbb{C}(X,X) = b\,Var(X) \Rightarrow a = \boxed{\frac{Cov(X,Y)}{Var(X)}}$

$$\text{et } \mathbb{E}(Y) = a\,\mathbb{E}(X) + b \Rightarrow \boxed{b = \mathbb{E}(Y) - a\,\mathbb{E}(X).}$$

- Cas général : on s'attend à trouver la même sol que le cas parfait. Cependant, les calculs sont plus compliqués, vu qu'on ne connaît pas e.

Soit $(X_1, Y_1), \dots, (X_n, Y_n)$ un échantillon iid. On $Y_i = aX_i + b + e_i$.
Le but est de miniser l'erreur. On utilise l'erreur quadratique :

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - aX_i - b)^2 \qquad (*)$$

Si a et b minisent l'erreur quadratique, alors en dérivant $(*)$, on obtient

(1) $\sum_{i=1}^{n} 2(Y_i - aX_i - b)(-X_i) = 0$ (en dérivant par rapport à a)

(2) $\sum_{i=1}^{n} 2(Y_i - aX_i - b)(-1) = 0$ (en dérivant par rapport à b)

$$(2) \Leftrightarrow 0 = \sum_{i=1}^{n} (Y_i - aX_i - b) = \sum_{i=1}^{n} Y_i - a\sum_{i=1}^{n} X_i - nb$$

$$\Leftrightarrow b = \frac{1}{n}\sum_{i=1}^{n} Y_i - a\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{Y} - a\bar{X}$$

$$(1)\ 0 = \sum_{i=1}^{n} (X_i Y_i - aX_i^2 - bX_i) = \sum_{i=1}^{n} X_i Y_i - a\sum_{i=1}^{n} X_i^2 - b\underbrace{\sum_{i=1}^{n} X_i}_{n\bar{X}}$$

$$= \sum_{i=1}^{n} X_i Y_i - a\sum_{i=1}^{n} X_i^2 - (\bar{Y} - a\bar{X})\, n\bar{X}$$

$$= \left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right) - a\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

$$= n\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}\right)}_{\text{Cov empirique}} - an\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)}_{\text{Variance empirique}}$$

$$\Rightarrow a = \frac{\widehat{\text{Cov}}(X,Y)}{\hat{\sigma}_X^2} \qquad \text{si } \hat{\sigma}_X^2 \neq 0$$

Conclusion : $a = \dfrac{\widehat{\mathrm{cov}}(X,Y)}{\hat{\sigma}_x^2} = \dfrac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} = \dfrac{\sum_{i=1}^n (x_i-\bar{x})^2 (Y_i-\bar{Y})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}$

$b = \bar{Y} - a\bar{X}.$

§ Distribution des estimateurs de a et b.

On peut voir a et b comme des paramètre et on peut utiliser les formules précédentes comme estimateurs :

$$\hat{a} = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\sigma_X^2} \qquad \hat{b} = \bar{Y} - \hat{a}\bar{X}.$$

On fait notre analyse dans le cas où on a observé les X_i, c'est-à-dire $X_i = x_i$, et l'on veut prédire Y_i à partir de x_i. On suppose également que $e_1, \dots, e_n$ sont iid avec distribution $N(0, \sigma^2)$.

Comme $Y_i = a x_i + b + e_i$, il suit que $Y_i \sim N(a x_i + b, \sigma^2)$.

Distribution a : $\hat{a} = \dfrac{\widehat{\mathrm{cov}}(X,Y)}{\hat{\sigma}_x^2} = \dfrac{\sum_{i=1}^n (x_i-\bar{x})(Y_i-\bar{Y})}{S_x^2}$

$= \sum_{i=1}^n \dfrac{(x_i-\bar{x})Y_i}{S_x^2}$

$\sum (x_i-\bar{x})\bar{Y} = (n\bar{x} - n\bar{x})\bar{Y} = 0$

où $S_x^2 = \sum (x_i-\bar{x})^2 = (n-1)\sigma_x^2$

C'est une somme de distributions normales indépendantes, donc $\hat{a}$ suit une loi normale. On a

$Y = ax + b + e$

$\mathbb{E}(Y) = ax + b$

$$\mathbb{E}(\hat{a}) = \sum_{i=1}^n \frac{(x_i-\bar{x})}{S_x^2} \left(\mathbb{E}(Y_i) - \mathbb{E}(\bar{Y})\right) = \sum_{i=1}^n \frac{(x_i-\bar{x})(ax_i + b - a\bar{x} - b)}{S_x^2}$$

$$= \frac{a}{S_x^2} \underbrace{\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})}_{S_x^2} = a$$

Donc $\hat{a}$ est sans biais.

Pour la variance (= EQM), on a

($Y_i \sim N(ax_i + b, \sigma^2)$)

$$\mathbb{V}ar(\hat{a}) = \frac{1}{s_x^4} \sum_{i=1}^{n} (x_i - \bar{x})^2 \mathbb{V}ar(Y_i) = \frac{\sigma^2}{s_x^4} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{= s_x^2} = \frac{\sigma^2}{s_x^2}.$$

Conclusion : $\hat{a} \sim N(a, \frac{\sigma^2}{s_x^2})$.

Distribution de $\hat{b}$: $\hat{b} = \bar{Y} - \hat{a}\bar{x}$

$\hat{b}$ est une combinaison linéaire des Y_i, donc $\hat{b}$ suit une loi normale.

$$\mathbb{E}(\hat{b}) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}(Y_k) - \mathbb{E}(\hat{a})\bar{x}$$

$$= b + \frac{a}{n} \sum_{k=1}^{n} x_i - a\bar{x}$$

$$= b$$

donc $\hat{b}$ est sans biais

$$\mathbb{V}ar(\hat{b}) = \mathbb{V}ar(\bar{Y}) + \bar{x}^2 \mathbb{V}ar(\hat{a})$$

$$= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{s_x^2}$$

$$= \sigma^2 \left(\frac{s_x^2 + n\bar{x}^2}{n s_x^2} \right)$$

$$= \frac{\sigma^2}{n s_x^2} \sum_{i=1}^{n} x_i^2$$

Donc $\hat{b} \sim N\left(b, \frac{\sigma^2}{n s_x^2} \sum_{i=1}^{n} x_i^2\right)$

De plus, on sait que $e_i = Y_i - aX_i - b$.

Estimation de la variance inconnue σ^2

On sait que les $e_i \sim N(0, \sigma^2)$. Ainsi, $\frac{e_i}{\sigma} \sim N(0,1)$, donc

$\sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} e_i^2$ suit une loi du χ^2. Il y aurait habituellement n

degrés de liberté, mais nous allons utiliser la relation $e_i = Y_i - aX_i - b$, où a et b sont

inconnues. Si on remplace un paramètre par son estimateur, on perd un degré de liberté (c'est un principe que nous accepterons). Dans notre, avec deux paramètres, on perd deux degrés de liberté, donc $\frac{1}{\sigma^2}\sum_{i=1}^{n} e_i \sim \chi^2_{n-2}$.

Il suit que $\mathbb{E}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} e_i\right) = n-2 \Rightarrow \mathbb{E}\left(\frac{1}{n-2}\sum_{i=1}^{n} e_i^2\right) = \sigma^2$.

Après des calculs fastidieux, on peut obtenir

$$\sum_{i=1}^{n} e_i^2 = \frac{S_x^2 S_Y^2 - S_{xY}^2}{S_x^2} = (n-1)\,\frac{\hat{\sigma}_x^2 \hat{\sigma}_Y^2 - \hat{\mathrm{cov}}(X,Y)^2}{\hat{\sigma}_x^2}$$

où $S_x^2 = (n-1)\hat{\sigma}_x^2$, $S_Y^2 = (n-1)\sigma_Y^2$ et $S_{xY} = (n-1)\hat{\mathrm{cov}}(X,Y)$

Coefficient de détermination : $R^2 = \frac{S_Y^2 - S_e^2}{S_Y^2} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_e^2}{\sigma_Y^2}$, où $S_Y^2 = (n-1)\hat{\sigma}_Y^2$, $S_e^2 = (n-1)\hat{\sigma}_e^2$

Par sa définition, on voit que

$$\sigma_Y^2 = R^2\sigma_Y^2 + \sigma_e^2$$

Variance expliquée par X — autre variance

$\sigma_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Heuristique : $Y = aX + b + \varepsilon$

$\Rightarrow \mathrm{Var}(Y) = a^2\mathrm{Var}(X) + \mathrm{Var}(\varepsilon)$

il y a une partie de la variance de Y qui est expliquée par la variance de X et une partie qui est expliquée par le fait que le modèle n'est pas parfait (le terme d'erreur)

Ainsi, R^2 est la proportion de la variance de Y qui est expliqué par la variance de X.

Propriété de R^2 : • $R^2 \in [0,1]$ (car $R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}$ et $\sigma_Y^2, \sigma_\varepsilon^2 \geq 0$)

• Si $R^2 = 0$, la variance de Y n'est pas linéairement expliquée par celle de X.

• Si $R^2 = 1$, la variance de Y est entièrement expliquée par celle de X

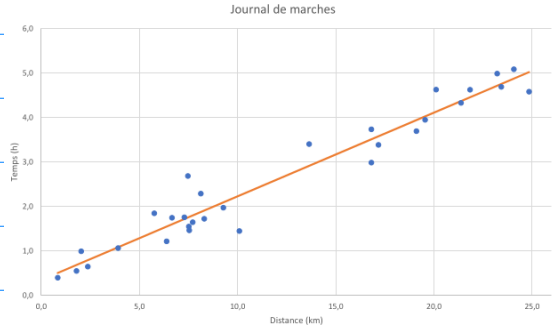
• $R = \sqrt{R^2}$ = corrélation = $\dfrac{\widehat{cov}(X,Y)}{\hat{\sigma}_X \, \sigma_Y}$

[En effet, on a $R^2 = \dfrac{\hat{\sigma}_Y^2 - \hat{\sigma}_e^2}{\hat{\sigma}_Y^2} = \dfrac{\hat{\sigma}_Y^2 - \sigma_X^2 \hat{\sigma}_Y^2 + \widehat{cov}(X,Y)^2 / \sigma_X^2}{\sigma_Y^2}$

$= \dfrac{\widehat{cov}(X,Y)^2}{\sigma_X^2 \sigma_Y^2} = cor(X,Y)^2$

où on a utilisé $\hat{\sigma}_e^2 = \dfrac{1}{(n-1)} \sum_{i=1}^{n} e_i^2 = (n-1) \dfrac{\hat{\sigma}_X^2 \hat{\sigma}_Y^2 - \widehat{cov}(X,Y)^2}{\hat{\sigma}_X^2}$]

Ex.



(pris dans les notes de cours d'Élise Dauignon)

Dans cette exemple, on trouve $R^2 = 0,944$, donc 94% de la variance dans le temps de marche est expliquée par la distance parcourue.

La droite en orange est la droite de régression $Y = \hat{a}X + \hat{b}$, où $\hat{a}$ et $\hat{b}$ sont calculées à partir des données en utilisant les formules de leur estimateur.

Inférence statistique sur a et b

On peut faire des tests d'hypothèse sur les valeurs de a et b.

P. ex. $H_0 : a = a_0$ $H_1 : a \neq a_0$.

Dans ce cas, si l'erreur suit une loi normale, alors $\hat{a} \sim N(a_0, \frac{\sigma^2}{S_x^2})$

On ne connaît pas la variance σ^2 de e, donc on utilise $\hat{\sigma}_e$:

$$T_{stat} = \frac{\hat{a} - a_0}{\hat{\sigma}_\varepsilon / (n-2)\, S_x} \sim \text{Student}(n-2).$$

Ex. On prend l'échantillon de données (distance (km), temps de marche (h)) = (X, Y) :

$(4, 1)$ $(7; 1,5)$ $(8; 1,6)$ $(17, 3)$

On s'intéresse à la régression $Y = aX + b$ et à l'hypothèse $H_0 : a = 0,2$.
On utilise $T_{stat} = \frac{\hat{a} - 0,2}{\hat{\sigma}_\varepsilon / (n-2) S_x}$. Pour calculer t_{stat}, il faut calculer

le $\hat{a}$ empirique, $\hat{\sigma}_\varepsilon^2$ et S_x^2. Pour calculer $\hat{\sigma}_\varepsilon^2$, il faut aussi calculer $\hat{b}$ empirique.

On rappelle que $\hat{a} = \frac{\widehat{cov}(X,Y)}{\sigma_x \sigma_Y}$ et $\hat{b} = \bar{Y} - \hat{a}\bar{X}$.

1. $\bar{x} = \frac{4 + 7 + 8 + 17}{4} = \frac{36}{4} = 9$ $\quad \bar{y} = 1,775$.

2. σ_x^2, σ_Y^2, $\widehat{cov}(X, Y)$:

i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-5	25	$-0,775$	0,6	3,875
2	-2	4	$-0,275$	0,075	0,55
3	-1	1	$-0,175$	0,306	0,175
4	8	64	1,225	1,5	9,8
$\sum$		94		2,2075	14,4

$\sigma_x^2 = \frac{94}{3} = 31,\overline{3}$ $\quad \sigma_Y^2 = \frac{2,2075}{3} = 0,735$ $\quad cov(X,Y) = \frac{14,4}{3} = 4,8$

$\sigma_x = 5,6$ $\quad \sigma_Y = 0,857$

3. $\hat{a} = \frac{4,8}{5,6 \times 0,857} = 1$ $\quad \hat{b} = 1,775 - 1 \times 9 = -7,225$

4. $e_i = y_i - \hat{a} x_i - b$

i	e_i	e_i^2
1	4,225	17,85
2	1,725	2,97
3	0,825	0,68
4	-6,775	45,9
Σ	X	67,4

$\hat{\sigma}_e^2 = \frac{67,4}{3} = 22,47$

$\hat{\sigma}_e = 4,74.$

5. $t_{stat} = \frac{\hat{a} - 0,2}{\hat{\sigma}_e / (n-2) S_x} = \frac{1 - 0,2}{4,74 / (2 \times 9,69)} = 3,27.$

$S_x^2 = (n-1) \hat{\sigma}_x^2$

$= 3 \times \hat{\sigma}_x^2 = 94$

$S_x = 9,69$

$dl = n - 2 = 4 - 2 = 2$

6. $T_{stat} \sim Student(2)$

valeur-p $= \mathbb{P}(|T_{stat}| \geq t_{stat} \mid H_0) = 0,082.$

7. Avec un seuil $\alpha = 5\%$, on a $8,2\% \geq \alpha$, donc on ne rejette pas H_0.