

Statistique Bivariée

Présenté par :
Pr. Abdelaziz Qaffou

EST-Beni Mellal

DUT: GI-SIR-IDIA-S3-2025-2026



Plan du cours

- 1 Organisation des données
- 2 Distributions marginales
- 3 Distributions conditionnelles
- 4 Liaison entre deux variables
- 5 Etude de deux variables qualitatives
- 6 Etude de deux variables quantitatives
 - Introduction
 - Régression linéaire simple
 - Estimation des paramètres
 - Corrélation



Outline

- 1 Organisation des données
- 2 Distributions marginales
- 3 Distributions conditionnelles
- 4 Liaison entre deux variables
- 5 Etude de deux variables qualitatives
- 6 Etude de deux variables quantitatives
 - Introduction
 - Régression linéaire simple
 - Estimation des paramètres
 - Corrélation



Exemple

Etude sur 5761 femmes de la survenue d'accouchement prématuré et de l'exposition à des événements stressants.

X: type d'accouchement (variable qualitative à 2 modalités)

Y: score sur une échelle allant de 0 à 3. Plus le score est élevé plus l'exposition à des événements stressants a été importante (variable quantitative discrète à 4 valeurs).

X\Y	0	1	2	3
à terme	4698	413	250	197
prématuré	165	16	12	10



Notations

On considère deux variables X et Y observées simultanément sur chacun des N individus de la population.

On notera x_i , $i = 1, \dots, k$ les k modalités ou valeurs de la variable X.

On notera y_j , $j = 1, \dots, l$ les l modalités ou valeurs de la variable Y.

On notera n_{ij} l'effectif correspondant au couple (x_i, y_j) .

Définition

On appelle distribution jointe des effectifs de X et Y l'ensemble des informations (x_i, y_j, n_{ij}) pour $i = 1, \dots, k$ et $j = 1, \dots, l$.



Tableau de contingence

On synthétise les données de la distribution jointe du couple (X, Y) par un tableau à double entrée appelé tableau de contingence:

$X \setminus Y$	y_1	...	y_j	...	y_l
x_1	n_{11}		n_{1j}		n_{1l}
...					
x_i	n_{i1}		n_{ij}		n_{il}
...					
x_k	n_{k1}		n_{kj}		n_{kl}



Exemple

$n_{23} = 12$: l'effectif des femmes qui ont accouché prématurément et qui ont un score égal à 2.

Remarque

$$\sum_{i=1}^k \sum_{j=1}^l n_{ij} = N$$



Proportions de la distribution jointe du couple (X, Y)

Exemple

$X \setminus Y$	0	1	2	3
à terme	4698	413	250	197
prématuré	165	16	12	10

Total $N = 5761$

$$\mathbb{P}(X = \text{à terme}, Y = 0) = \frac{4698}{5761} = 0.815.$$

$X \setminus Y$	0	1	2	3
à terme	0.815	0.072	0.043	0.034
prématuré	0.029	0.003	0.002	0.002



Définition

La proportion du couple (x_i, y_j) est $\mathbb{P}(X = x_i, Y = y_j) = p_{ij} = \frac{n_{ij}}{N}$.

Remarque

$$\sum_{i=1}^k \sum_{j=1}^l p_{ij} = 1$$



Outline

- 1 Organisation des données
- 2 Distributions marginales
- 3 Distributions conditionnelles
- 4 Liaison entre deux variables
- 5 Etude de deux variables qualitatives
- 6 Etude de deux variables quantitatives
 - Introduction
 - Régression linéaire simple
 - Estimation des paramètres
 - Corrélation



On ajoute au tableau de contingence les totaux en ligne et en colonne.

X\Y	y_1	...	y_j	...	y_l	Totaux
x_1	n_{11}		n_{1j}		n_{1l}	$n_{1\bullet}$
...						
x_i	n_{i1}		n_{ij}		n_{il}	$n_{i\bullet}$
...						
x_k	n_{k1}		n_{kj}		n_{kl}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet l}$	N



Exemple

X\Y	0	1	2	3	Totaux
à terme	4698	413	250	197	$n_{1\bullet} = 5558$
prématuré	165	16	12	10	$n_{2\bullet} = 203$
Totaux	4863	429	262	207	5761
	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	$n_{\bullet 4}$	



En marge à droite (totaux en ligne) :

X	
à terme	$n_{1\bullet} = 5558$
prématuré	$n_{2\bullet} = 203$
Total	N=5761

Distribution marginale de X:

X	à terme	prématuré	Total
Effectif $n_{i\bullet}$	5558	203	5761



Pour chaque indice i , l'effectif $n_{i\bullet}$ est l'effectif de la modalité x_i de X quelle que soit la modalité de Y.

$$n_{i\bullet} = \sum_{j=1}^k n_{ij} = \text{total de la ligne } i.$$

Les k couples $(x_i, n_{i\bullet})$ définissent la **distribution marginale** de la variable X.

Remarque

$$\sum_{i=1}^k n_{i\bullet} = N$$



Proportions marginales pour X

X	à terme	prématuré	Total
Effectif $n_{i\bullet}$	5558	203	N=5761
Proportion $p_{i\bullet}$	0.964	0.036	1

Définition

La proportion marginale de x_i est

$$\mathbb{P}(X = x_i) = p_{i\bullet} = \frac{n_{i\bullet}}{N}$$



En marge en bas (totaux en colonne):

X\Y	0	1	2	3	Totaux
à terme	4698	413	250	197	5558
prématuré	165	16	12	10	203
Totaux	4863	429	262	207	5761

Distribution marginale de Y:

Y	0	1	2	3	Total
Effectif $n_{\bullet j}$	4863	429	262	207	N=5761



Pour chaque indice j , l'effectif $n_{\bullet j}$ est le nombre total d'observations de la modalité y_j quelle que soit la modalité de X.

$$n_{\bullet j} = \sum_{i=1}^k n_{ij} = \text{total de la colonne } j$$

Les I couples $(y_j, n_{\bullet j})$ définissent la **distribution marginale** de la variable Y

Remarque

$$\sum_{j=1}^I n_{\bullet j} = N$$



Proportions marginales pour Y

Y	0	1	2	3	Total
Effectif $n_{\bullet j}$	4863	429	262	207	5761
Proportion $p_{\bullet j}$	0.844	0.074	0.045	0.036	1

Définition

La proportion marginale de y_j est

$$\mathbb{P}(Y = y_j) = p_{\bullet j} = \frac{n_{\bullet j}}{N}$$



Outline

- 1 Organisation des données
- 2 Distributions marginales
- 3 **Distributions conditionnelles**
- 4 Liaison entre deux variables
- 5 Etude de deux variables qualitatives
- 6 Etude de deux variables quantitatives
 - Introduction
 - Régression linéaire simple
 - Estimation des paramètres
 - Corrélation



Le principe des distributions conditionnelles est de décrire le comportement de l'une des deux variables quand l'autre a une valeur donnée.

Exemple

X\Y	0	1	2	3	Totaux
à terme	4698	413	250	197	5558
prématuré	165	16	12	10	203
Totaux	4863	429	262	207	5761



Ligne 2 du tableau de contingence : distribution de la variable Y chez les femmes ayant eu un accouchement prématuré.

Exemple

$Y X = \text{prématuré}$	0	1	2	3	Totaux
Effectif	165	16	12	10	$n_{2\bullet} = 203$
Proportion	0.813	0.079	0.059	0.049	1

Ce tableau donne la distribution conditionnelle de Y sachant que l'accouchement a été prématuré.



Exemple

X\Y	0	1	2	3	Totaux
à terme	4698	413	250	197	5558
prématuré	165	16	12	10	203
Totaux	4863	429	262	207	5761

Colonne 3 du tableau de contingence: distribution conditionnelle de X sachant que la femme enceinte a subi un stress de niveau 2:

X $Y=2$	à terme	prématuré	Total
Effectif	250	12	$n_{\bullet 3} = 262$
Proportion	0.954	0.046	1



A la ligne i du tableau de contingence, on lit la distribution conditionnelle de la variable Y sachant que la variable X prend la modalité x_i ; elle est notée distribution conditionnelle de $Y | X=x_i$.

Il y a k distributions conditionnelles de Y sachant $X = x_i$.

A la colonne j du tableau de contingence, on lit la distribution conditionnelle de la variable X sachant que la variable Y prend la modalité y_j ; elle est notée distribution conditionnelle de $X | Y=y_j$.

Il y a l distributions conditionnelles de X sachant $Y = y_j$.



Outline

- 1 Organisation des données
- 2 Distributions marginales
- 3 Distributions conditionnelles
- 4 Liaison entre deux variables
- 5 Etude de deux variables qualitatives
- 6 Etude de deux variables quantitatives
 - Introduction
 - Régression linéaire simple
 - Estimation des paramètres
 - Corrélation



Exemple

Distribution marginale de Y:

Y	0	1	2	3	Total
Proportion $p_{\bullet j}$	0.844	0.074	0.045	0.036	1

Distribution conditionnelle de Y sachant que l'accouchement a été prématuré:

$Y X = \text{prématué}$	0	1	2	3	Total
Proportion	0.813	0.079	0.059	0.049	1

Distribution conditionnelle de Y sachant que l'accouchement était à terme:

$Y X = \text{à terme}$	0	1	2	3	Total
Proportion	0.845	0.074	0.045	0.035	1



Autre exemple

X\Y	0	1	2	3	Totaux
à terme	4140	862	278	278	5558
prématuré	22	30	51	100	203
Totaux	4162	892	329	378	5761

Distribution marginale de Y:

Y	0	1	2	3	Total
Proportion $p_{\bullet j}$	0.722	0.155	0.057	0.066	1

Distribution conditionnelle de Y sachant que l'accouchement a été prématuré:

$Y X = \text{prématuré}$	0	1	2	3	Total
Proportion	0.108	0.148	0.251	0.493	1



D

istribution conditionnelle de Y sachant que l'accouchement était à terme:

$Y X = \text{à terme}$	0	1	2	3	Total
Proportion	0.745	0.155	0.050	0.050	1



On peut comparer les distributions conditionnelles de Y sachant X entre elles et à la distribution marginale de Y.

Si ces distributions sont très proches, on peut conjecturer une certaine indépendance entre les deux variables.

Si ces distributions sont très distinctes, cela signifie que les modalités de X ont une influence sur la variable Y et donc que les deux variables sont liées.

Remarque

de même , on peut comparer les distributions conditionnelles de X sachant Y à la distribution marginale de X.



Outline

- 1 Organisation des données
- 2 Distributions marginales
- 3 Distributions conditionnelles
- 4 Liaison entre deux variables
- 5 **Etude de deux variables qualitatives**
- 6 Etude de deux variables quantitatives
 - Introduction
 - Régression linéaire simple
 - Estimation des paramètres
 - Corrélation



Données observées

Si les deux variables x et y sont qualitatives, alors les données observées sont une suite de couples de variables:

$$(x_1, y_1), \dots, (x_i, y_j), \dots, (x_n, y_n),$$

chacune des deux variables prend comme valeurs des modalités qualitatives.

Les valeurs distinctes de x et y sont notées respectivement:

$$x_1, \dots, x_j, \dots, x_J$$

et

$$y_1, \dots, y_k, \dots, y_K.$$



Tableau de contingence

Les données observées peuvent être regroupées sous la forme d'un tableau de contingence

$X \setminus Y$	y_1	...	y_k	...	y_K	Total
x_1	n_{11}	...	n_{1k}	...	n_{1K}	$n_{1\bullet}$
...
x_j	n_{j1}	...	n_{jk}	...	n_{jK}	$n_{i\bullet}$
...
x_J	n_{J1}	...	n_{Jk}	...	n_{JK}	$n_{J\bullet}$
Total	$n_{\bullet 1}$...	$n_{\bullet k}$...	$n_{\bullet K}$	n



Tableau de contingence

Les $n_{j\cdot}$ et $n_{\cdot k}$ sont appelés les effectifs marginaux. Dans ce tableau,

- $n_{j\cdot}$ représente le nombre de fois que la modalité x_j apparaît,
- $n_{\cdot k}$ représente le nombre de fois que la modalité y_k apparaît,
- n_{jk} représente le nombre de fois que les modalités x_j et y_k apparaissent ensemble.

On a les relations

$$\sum_{j=1}^J n_{jk} = n_{\cdot k}, \text{ pour tout } k = 1, \dots, K,$$

$$\sum_{k=1}^K n_{jk} = n_{j\cdot}, \text{ pour tout } j = 1, \dots, J,$$

et

$$\sum_{j=1}^J n_{j\cdot} = \sum_{k=1}^K n_{\cdot k} = \sum_{j=1}^J \sum_{k=1}^K n_{jk} = n.$$



Exemple

On s'intéresse à une éventuelle relation entre le sexe de 200 personnes et la couleur des yeux. Le Tableau suivant reprend le tableau de contingence:

	Bleu	Vert	Maron	Total
Homme	10	50	20	80
Femme	20	60	40	120
Total	30	110	60	200



Tableau des fréquences

Le tableau de fréquences s'obtient en divisant tous les effectifs par la taille de l'échantillon:

$$f_{jk} = \frac{n_{jk}}{n}, j = 1, \dots, J, k = 1, \dots, K$$

$$f_{j\cdot} = \frac{n_{j\cdot}}{n}, j = 1, \dots, J, f_{\cdot k} = \frac{n_{\cdot k}}{n}, k = 1, \dots, K.$$

Le tableau des fréquences est:

X\Y	y_1	...	y_k	...	y_K	Total
x_1	f_{11}	...	f_{1k}	...	f_{1K}	$f_{1\bullet}$
...
x_j	f_{j1}	...	f_{jk}	...	f_{jK}	$f_{i\bullet}$
...
x_J	f_{J1}	...	f_{Jk}	...	f_{JK}	$f_{k\bullet}$
Total	$f_{\bullet 1}$...	$f_{\bullet k}$...	$f_{\bullet K}$	1



Exemple

On reprend le tableau de contingence précédent:

	Bleu	Vert	Maron	Total
Homme	0.05	0.25	0.10	0.40
Femme	0.10	0.30	0.20	0.60
Total	0.15	0.55	0.30	1.00



Profils lignes et profils colonnes

Un tableau de contingence s'interprète toujours en comparant des fréquences en lignes ou des fréquences en colonnes (appelés aussi profils lignes et profils colonnes).

Les profils lignes sont définis par

$$f_k^j = \frac{n_{jk}}{n_j} = \frac{f_{jk}}{f_{j\cdot}}, \quad k = 1, \dots, K, j = 1, \dots, J,$$

et les profils colonnes par

$$f_j^k = \frac{n_{jk}}{n_{\cdot k}} = \frac{f_{jk}}{f_{\cdot k}}, \quad k = 1, \dots, K, j = 1, \dots, J.$$



Exemple

On reprend le même tableau de contingence précédent, Les profils lignes sont:

	Bleu	Vert	Maron	Total
Homme	0.13	0.63	0.25	1.00
Femme	0.17	0.50	0.33	1.00
Total	0.15	0.55	0.30	1.00

Les profils colonnes sont:

	Bleu	Vert	Maron	Total
Homme	0.33	0.45	0.33	0.40
Femme	0.67	0.55	0.67	0.60
Total	1.00	1.00	1.00	1.00



Effectifs théoriques et khi-carré

On cherche souvent une interaction entre des lignes et des colonnes, un lien entre les variables. Pour mettre en évidence ce lien, on construit un tableau d'effectifs théoriques qui représente la situation où les variables ne sont pas liées (indépendance). Ces effectifs théoriques sont construits de la manière suivante:

$$n_{jk}^* = \frac{n_j \cdot n_k}{n}.$$

Les effectifs observés n_{jk} ont les mêmes marges que les effectifs théoriques n_{jk}^* .

Enfin, les écarts à l'indépendance sont définis par

$$e_{jk} = n_{jk} - n_{jk}^*.$$



Effectifs théoriques et khi-carré

La dépendance du tableau se mesure au moyen du khi-carré défini par

$$\chi^2_{obs} = \sum_{k=1}^K \sum_{j=1}^J \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}.$$

On pose l'hypothèse nulle:

H_0 : Il n'y a pas de relation entre les deux variables (les deux variables sont indépendantes).

La règle de décision: On rejette l'hypothèse nulle si $\chi^2_{obs} > \chi^2_{0,05; ddl}$ avec ddl: degré de liberté=(nombre de lignes-1)(nombre de colonne-1).



V de Cramer

On peut utiliser aussi le V de Cramer est défini par

$$V = \sqrt{\frac{\chi^2_{obs}}{n \min(J-1, K-1)}}.$$

Le V de Cramer est compris entre 0 et 1. Il ne dépend ni de la taille de l'échantillon ni de la taille du tableau. Si $V \approx 0$, les deux variables sont indépendantes. Si $V \approx 1$, il existe une relation fonctionnelle entre les variables, ce qui signifie que chaque ligne et chaque colonne du tableau de contingence ne contiennent qu'un seul effectif différent de 0 (il faut que le tableau ait le même nombre de lignes que de colonnes).



Outline

- 1 Organisation des données
- 2 Distributions marginales
- 3 Distributions conditionnelles
- 4 Liaison entre deux variables
- 5 Etude de deux variables qualitatives
- 6 Etude de deux variables quantitatives
 - Introduction
 - Régression linéaire simple
 - Estimation des paramètres
 - Corrélation



But

Etablir un lien entre une variable dépendante Y et une variable indépendante X pour pouvoir ensuite faire des prévisions sur Y lorsque X est mesurée.

Exemple 1

L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement du produit a donné les valeurs suivantes pour la température x_i et le rendement correspondant y_i :

Température °C	Rendement %	Température °C	Rendement %
100	45	150	70
110	51	160	74
120	54	170	78
130	61	180	85
140	66	190	89



Le graphe ci-dessous représente les points $(x_i; y_i)$ pour ces données et suggère une relation linéaire entre X et Y.

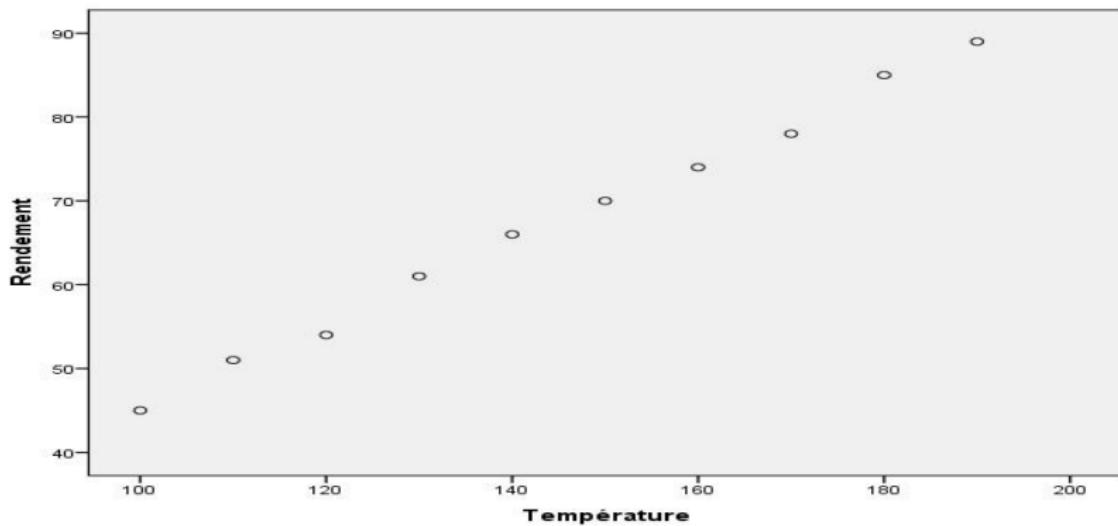


Figure 1: Diagramme de dispersion entre la température et le rendement de l'exemple 1



Définition du modèle

Un modèle de régression linéaire simple est de la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

où

- Y est la variable dépendante (variable à expliquer).
- β_0 et β_1 sont les coefficients (ordonnée à l'origine et pente), deux paramètres à estimer.
- X est la variable indépendante (variable explicative).
- ε est le terme d'erreur aléatoire du modèle.



Hypothèses sur le modèle

Pour n observations, on peut écrire le modèle de régression linéaire simple sous la forme:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

On suppose que:

- ε_i est une variable aléatoire, non observée,
- x_i est observée et non aléatoire,
- y_i est observée et aléatoire.

On fait les trois hypothèses additionnelles suivantes:

Hypothèse 1 (les erreurs sont centrées):

$$\mathbb{E}[\varepsilon_i] = 0, \quad \forall i = 1, \dots, n,$$

ou de manière équivalente:

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i, \quad \forall i = 1, \dots, n,$$



Hypothèses sur le modèle (suite)

Hypothèse 2 (homoscédasticité):

$$\mathbb{V}[\varepsilon_i] = \sigma^2, \quad \forall i = 1, \dots, n,$$

ou de manière équivalente:

$$\mathbb{V}[y_i] = \sigma^2, \quad \forall i = 1, \dots, n,$$

Cette variance σ^2 est un paramètre du modèle qu'il faudra estimer.



Hypothèses sur le modèle (suite)

Hypothèse 3 (indépendance des erreurs):

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j,$$

ou de manière équivalente:

$$\text{Cov}(y_i, y_j) = 0, \quad \forall i \neq j,$$

Sous cette hypothèse, les termes d'erreur ε_i sont non corrélés (indépendantes).



Estimation des paramètres β_0 et β_1 .

Supposons que n paires d'observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ont été faites. Substituant dans le modèle linéaire, on obtient

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow \varepsilon_i = y_i - \beta_0 - \beta_1 x_i.$$

Les coefficients sont déterminés par la méthode des moindres carrés qui minimise la somme des carrés des erreurs:

$$L(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Le minimum est atteint pour $\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = 0$ et $\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = 0$
soit après quelques calculs:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ et } \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$



Estimation du paramètre σ^2

$\hat{\beta}_0$ et $\hat{\beta}_1$ sont les estimateurs de β_0 et β_1 successivement qui minimisent la somme des carrés des résidus

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

où \hat{y}_i est valeur prédite par le modèle lorsque $x = x_i$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Le minimum de $L(\beta_0, \beta_1)$ est égal à $\sum_{i=1}^n \hat{\varepsilon}_i^2$. On note:

la somme des carrés des résidus (SCR) par $SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$.

la somme des carrés dûe à la regression (SCE) par $SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

la somme des carrés totale (SCT) par $SCT = SCE + SCR$.



Estimation du paramètre σ^2 (suite)

Le paramètre σ^2 est défini par

$$\sigma^2 = \mathbb{V}(\varepsilon_i) = \mathbb{V}(y_i) = \mathbb{E}[(y_i - \mathbb{E}[y_i])^2].$$

En prenant $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ comme estimateur de $\mathbb{E}[y_i]$, il apparaît naturel d'estimer σ^2 par

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

La perte de deux degrés de liberté dans l'expression de s^2 est le "coût" de l'estimation de β_0 et de β_1 nécessaire pour obtenir les \hat{y}_i .



Coefficient de détermination

Le coefficient de détermination du modèle de régression linéaire est

$$R^2 = \frac{SCE}{SCT}$$

Le coefficient R^2 mesure le pourcentage de la variabilité totale SCT qui est expliquée par le modèle.

Si R^2 est proche de 1, alors le modèle semble adéquat.



Coefficient de corrélation

La corrélation entre deux variables aléatoires X et Y est mesurée par le coefficient

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

Le coefficient de corrélation échantillonnaux est

$$r = \frac{S_{XY}}{S_{XX} S_{YY}}.$$

Le coefficient de corrélation ρ est estimé ponctuellement par r .



Interprétation du coefficient de corrélation

On peut montrer que $-1 \leq r \leq 1$.

- Si $r = -1$ ou $r = 1$ alors il y a corrélation parfaite entre X et Y et les points (x_i, y_i) sont tous sur la droite de régression.
- Si $r = 0$ alors il n'y a pas de corrélation entre X et Y et les points (x_i, y_i) sont dispersés au hasard.
- Si $0 < r < 1$ alors il y a corrélation positive faible, moyenne ou forte entre X et Y . Dans ce cas, une augmentation de X entraîne une augmentation de Y .
- Si $-1 < r < 0$ alors il y a corrélation négative faible, moyenne ou forte entre X et Y . Dans ce cas, une augmentation de X entraîne une diminution de Y .



Merci pour votre attention

