

CORRELATIONS ET REGRESSIONS

Relations entre deux ou plusieurs variables observée sur le même échantillon

1

PLAN du cours

- Evaluation graphique de la relation entre deux variables quantitatifs
- Coefficients de corrélation Pearson, Spearman
- Régression linéaire simple et multiple

2

2

Formulation du problème

Etude de la **relation / lien / association** entre **2 variables quantitatives**:

- Le poids et la tension artérielle systolique?
 - La taille du cerveau et le niveau d'intelligence?
- Poids = X: X_1, X_2, \dots, X_n
TAS = Y: Y_1, Y_2, \dots, Y_n

0. L'évaluation **graphique** de la relation :

- La diagramme de dispersion/nuage des points/scatter plot

1. L'**existence** d'une relation entre les variables X et Y

- A l'aide d'un test statistique sur le coefficient de corrélation

2. L'**intensité** / la force/l'importance/le degré de cette relation et la direction

- Le coefficient de corrélation Pearson ou Spearman
- Le coefficient de détermination
- La pente (coefficient b_1 des variables indépendantes) du régression

3. **Prédiction** : prédire les valeurs d'une variable sachant les valeurs de l'autre

- La régression
 - déterminer une fonction (fonction de régression) telle que $Y=f(X)$?

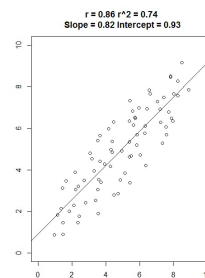
3

3

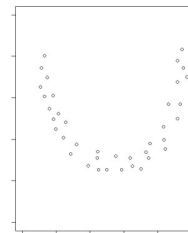
Evaluation graphique : diagramme de dispersion (nuage des points/scatter)

1^{er} objectif - Evaluer la linéarité

Si le nuage des points semble suggérer qu'il y a une tendance que les points sont plutôt situés autour d'une droite imaginaire – la relation est peut être linéaire



Si le nuage des points semble suggérer des tendances qui ne sont pas linéaires, la relation est peut être non linéaire (exponentielle, quadratique, ...)



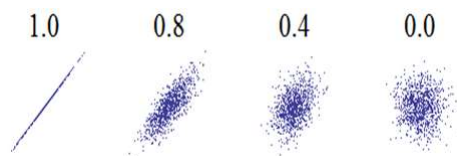
4

4

Evaluation graphique : diagramme de dispersion (nuage des points/scatter)

2eme objectif - Evaluer l'importance/la force/puissance de la corrélation/relation/lien/association:

Si la relation est plus probable linéaire, on peut évaluer d'une manière subjective la force/puissance/importance de la corrélation linéaire. Plus les points se rapprochent d'une droite de tendance, plus la corrélation/relation/lien/association est forte, plus les points sont distants, plus la corrélation est faible



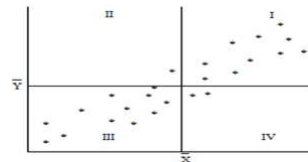
5

5

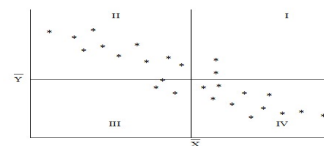
Evaluation graphique : diagramme de dispersion (nuage des points/scatter)

3eme objectif : évaluation visuelle de la relation entre deux variables quantitatifs en utilisant des cadrans (créés par les moyennes du x et du y) pour identifier la **tendance/sens/direction** directe/inversement proportionnelle, ou purement une évaluation visuelle de la tendance:

i) La plupart des points sont dans les cadrans I et III OU une sensation visuelle que les points ont une tendance croissante (plus la valeur du x augmente le y augmente aussi) \Rightarrow tendance croissante/ pente ascendante/ pente positive/ lien (direct) proportionnel



ii) La plupart des points sont dans les cadrans II et IV OU une sensation visuelle que les points ont une tendance décroissante (plus la valeur du x augmente le y diminuent) \Rightarrow tendance décroissante / pente descendante/ pente négative/ lien inversement proportionnelle



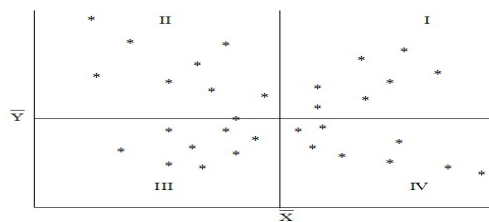
6

6

Evaluation graphique : diagramme de dispersion (nuage des points/scatter)

3eme objectif : évaluation visuelle de la relation entre deux variables quantitatives en utilisant des cadrans (créés par les moyennes du x et du y) pour identifier **la tendance/sens/direction** directe/inversement proportionnelle, ou purement une évaluation visuelle de la tendance (continuation):

- iii. les points sont distribués uniformément dans tous les cadrans ou les points semblent distribués au hasard partout \Rightarrow aucune tendance



7

7

Le coefficient de Corrélation linéaire Pearson

But: évaluer l'association des deux variables quantitatives du point de vue de la **direction** de l'association et l'**importance (force, degré)** de l'association

Le **coefficient de corrélation Pearson**:

- est une mesure de la **corrélation linéaire (force de l'association linéaire)** des deux variables quantitatives
- montre le dégré de rapprochement des points à une droite qui passe entre les points.

Condition d'application:

les paires des observations indépendantes dans l'échantillon
variables quantitatives

les deux variables soient normalement distribuées

(les variables suivent une distribution normale bi variée)

la relation entre les deux variables est linéaire simple (n'est pas quadratique, exponentielle...)

8

Indice de corrélation linéaire Pearson

Le calcul du coefficient de corrélation linéaire Pearson

Covariance d'échantillonnage $COV(X,Y)$:
$$COV(X,Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Coefficient de corrélation linéaire Pearson
$$r = \frac{COV(X,Y)}{S_X \cdot S_Y}$$

Coefficient de détermination $d = r * r$

X_i, Y_i – sont les valeurs des deux séries des données. \bar{X} et \bar{Y} sont les moyennes des deux séries. n – nombre des observations. S_X et S_Y sont les déviations standard d'échantillonnage, r – coefficient de corrélation Pearson, d – coefficient de détermination

9

9

Corrélation linéaire - interprétations

Covariance $COV(X,Y)$:

- > 0 tendance croissante/ pente ascendante/ lien direct proportionnel/ covariance positive
- < 0 tendance décroissante/ pente descendante/ lien inversement proportionnel/ covariance négative
- $\cong 0 \Rightarrow$ aucune tendance

r (coefficient de corrélation Pearson):

montre la **direction** et l'**intensité** de la corrélation;

Interprétation du direction/sens/tendance:

- > 0 tendance croissante/ pente ascendante/ lien direct proportionnel/ corrélation positive
- < 0 tendance décroissante/ pente descendante/ lien inversement proportionnel/ corrélation négative
- $\cong 0 \Rightarrow$ aucune tendance
- plus **r** ou la covariance est grand (en valeur absolue) plus la relation est forte / intense / puissante / importante
- plus **r** ou la covariance est proche du 0, plus la relation est faible / moins intense / moins puissante / moins importante

10

10

Le coefficient Pearson - interprétation

Interprétation de l'intensité/force/degré/importance de la corrélation linéaire avec les règles empiriques de Colton [Colton T. *Statistics in Medicine*. Little Brown and Company, New York, NY 1974] (**on préfère le mot corrélation ici**, même si association/liens/relation peut être utilisé)

[0 et 0.25) ou (-0.25 et 0] => une relation **négligeable** ou **aucune** corrélation linéaire entre les variables

[0.25 et 0.50) ou [-0.25 et -0.50) => un degré de corrélation **faible/acceptable**

[0.50 et 0.75) ou [-0.50 et -0.75) => un degré de corrélation **modérée à bonne**

[0.75 et 1] ou [-0.75 et -1] => une **très bonne à excellente** corrélation

Il y a autre divisions possibles aussi.

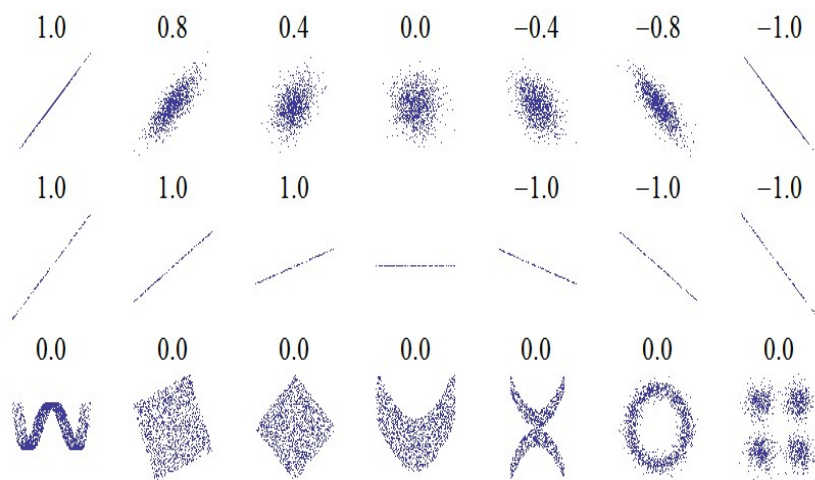
Ces règles doit être utilisée avec soins. Elle sont pour donner une idée, mais pour chaque problème, l'intensité de la relation est relative au domaine. Pour certain situations les valeurs en dessous de 0,8 peut être faibles.

(notation – les parenthèses [ou] indique que la valeur est inclus, et les parenthèses (ou) indiquent que la valeur n'est pas incluse dans l' intervalle)

11

11

Relation entre le coefficient de corrélation et différents types de nouages des points



12

Wikimedia commons

12

Calcul du coef. de corrélation Pearson

- Les valeurs du HDL Cholestérol (mmol/l) et les poids (kg) pour 10 malades sont (quantitative, paires des observations indépendantes, normale distribuées. La relation est linéaire)
- On a les moyennes de cholestérol et de poids $\bar{x} = 5.24, \bar{y} = 74.9$ et la variance de cholestérol = 1,15, la variance de poids = 214,89 et par conséquence les écarts types $s_x = 1,07, s_y = 14,65$.
- La covariance sur la population est donnée par la formule:
- $COV(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- $COV(x,y) = (1/9) \times [(6.8-5.24) \times (90-74.9) + (5.3-5.24) \times (75-74.9) + \dots]$

$$r = \frac{COV(x,y)}{s_x \times s_y} = \frac{14.324}{1.07 \times 14.65} = 0.91$$

Interprétation :

$r > 0 \Rightarrow$ tendance croissante / relation (directe) proportionnelle, $r > 0,75 \Rightarrow$ intensité de corrélation très bonne (règles du Colton)

HDL	6.8	5.3	4.3	5.0	7.1	5.5	3.8	4.6	4.0	6.0
Poids	90	75	70	73	110	67	60	65	59	80

13

13

Test statistique de signifiante pour le coefficient de corrélation linéaire du Pearson

- **But:** tester si deux variables ont une **corrélation linéaire** statistiquement significative / tester si deux variables **quantitatifs** sont **linéaire associées** d' une manière statistiquement significatif
- On utilise un test statistique sur le coefficient de corrélation pour voir s' il est différent de 0 (H_0 = absence de la corrélation)
- **Condition d' application:**
 - les paires des observations indépendantes dans l'échantillon
 - variables quantitatifs
 - les deux variables soient normalement distribuées
 - (les variables suit un distribution normale bi variée)
 - homoscédasticité - variance constante des erreurs (résidus)
 - sans valeurs aberrantes (très éloignes du nuage des points)

14

Test statistique de signifiante pour le coefficient de corrélation linéaire du Pearson 15

- L'hypothèse nulle:
 - il n'y a pas d'une différence statistiquement significative entre le coefficient de corrélation (des variables x et y) et 0
 - il n'y a pas d'une corrélation linéaire statistiquement significative entre les variables x et y
- L'hypothèse alternative:
 - il y a une différence statistiquement significative entre le coefficient de corrélation (des variables x et y) et 0
 - il y a une corrélation linéaire statistiquement significative entre les variables x et y
- La décision avec p-value.
 - Si $p\text{-value} < \alpha (=0,05) \Rightarrow$ on rejette H_0 et on accepte H_1
 - Si $p\text{-value} \geq \alpha (=0,05) \Rightarrow$ on ne peut pas rejeter H_0

15

Test statistique de signifiante pour le coefficient de corrélation linéaire du Pearson 16

Exemple

Le coefficient de corrélation linéaire Pearson pour la relation entre les triglycérides et le poids pour 50 sujets est 0,72, et la valeur du p associée est 0,001. Les paires des observations sont indépendantes, les données sont normalement distribuées, la relation est linéaire

Interprétation de la valeur du p associée au coefficient de corrélation:

$\Rightarrow p < 0,05$ il y a une **corrélation linéaire statistiquement significative entre les triglycérides et le poids**

La **corrélation linéaire** entre les **triglycérides** et le **poids** est **statistiquement significative**

Interprétation de la direction et de l'intensité de la corrélation: la relation est (directe) proportionnelle ($r=0,72 > 0$), et l' intensité de la corrélation est modérée à bonne ($r=0,72$ – est dans $[0,50$ et $0,75)$)

16

Test statistique de significiance pour le coefficient de 17 corrélacion linéaire du Pearson

Exemple

Le **coefficient de corrélation linéaire Pearson** pour la relation entre les triglycérides et la longueur du cheveux pour 24 sujets est 0,39, et la **valeur du p** associée est 0,35. Les paires des observations sont indépendantes, les données sont normalement distribuées, la relation est linéaire

Interprétation de la valeur du p associée au coefficient de corrélation:

=> $p > 0,05$ on ne peut pas dire qu'il y a une **corrélacion linéaire statistiquement significative** entre les triglycérides et la longueur du cheveux

Interprétation de la direction et de l'intensité de la corrélation:

l'interprétation n'a aucune utilité parce que la relation n'a pas atteint la signification statistique – donc le résultat peut être beaucoup influencée par la chance. Pour le moment on sait rien. La relation peut sembler directe proportionnelle ($r = 0,35 > 0$), et l'intensité de la corrélation pourrait être faible/acceptable ($r = 0,35$ – est dans $[0,25 \text{ et } 0,50]$) – si la relation était statistiquement significative. Mais il n'est pas ... donc on ne peut rien dire.

17

Coefficient de corrélation Spearman

But: évaluer l'association des deux variables du point de vue de la direction de l'association et l'importance de l'association

Condition d'application:

- Les paires des observations indépendantes dans l'échantillon
- Deux variables **quantitatives** (au moins une **non normalement** distribuées)
- Ou Les deux variables sont **ordinales**/
- Ou Une variable **ordinaire** et une quantitative (normalement ou non normalement distribuées)

Utilité: évaluer la relation entre

- deux variables quantitatives qui ne sont pas normalement distribuées
- deux variables quantitatives: une variable normale distribuée, et une autre non normale distribuée
- deux variables ordinales
- une ordinaire et une quantitative (n'importe s'il est normale distribuée, ou non)

18

18

Coefficient de corrélation de Spearman

Etapes de calcul:

- ❑ Remplacer la série bi variée $(x_1, \dots, x_n; y_1, \dots, y_n)$
 - avec la série des rangs $(R_{x1}, \dots, R_{xn}; R_{y1}, \dots, R_{yn})$, des valeurs x_i et y_i après leurs rangement dans ordre croissante
 - pour les valeurs égaux on prend la moyenne arithmétique des rangs.
- ❑ Calculer le coefficient ρ de Spearman:

Le coefficient r_s (ou rho):

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \text{ ou } d_i = R_{x_i} - R_{y_i}$$

On ne doit pas savoir le calculer pour l'examen, seulement l'interpréter, savoir quand l'utiliser

19

19

Coefficient de corrélation Spearman

Interprétation:

- $> 0 \Rightarrow$ tendance croissante/ascendant – lien (direct) proportionnel - positif
- $< 0 \Rightarrow$ tendance décroissante/descendant – lien inversement proportionnel - négatif
- $\cong 0 \Rightarrow$ aucune tendance
- plus r_s est grand (en valeur absolue) plus la relation est forte/ intense / puissante / importante
- plus r_s est proche du 0, plus la relation est faible / moins intense / moins puissante / moins importante

20

20

Coefficient de corrélation Spearman

Il y a aussi un test statistique pour vérifier si le coefficient de corrélation Spearman est statistiquement différent de 0.

- L'hypothèse nulle:

- il n'y a pas d'une différence statistiquement significative entre le coefficient de corrélation (des variables x et y) et 0
- il n'y a pas d'une corrélation statistiquement significative entre les variables x et y

- L'hypothèse alternative:

- il y a une différence statistiquement significative entre le coefficient de corrélation (des variables x et y) et 0
- il y a une corrélation statistiquement significative entre les variables x et y

- La décision avec p-value.

- Si $p\text{-value} < \alpha (=0,05) \Rightarrow$ on rejette H_0 et on accepte H_1
- Si $p\text{-value} \geq \alpha (=0,05) \Rightarrow$ on ne peut pas rejeter H_0

21

21

Régression

- **But:** - méthode pour étudier les relations entre 2 ou plusieurs variables

- **Forme générale:**

$$Y = f(X) + \varepsilon \text{ ou } f = \text{fonction de régression}$$

Concepts utilisés:

X = variable **indépendante**, explicative, prédictive
(en anglais: independent, explanatory, predictive)

Y = variable **dépendante**, expliquée, prédite
(en anglais: dependent, explained, predicted)

ε = écart (erreur) de cette approximation

22

Régression

Type de régression: Selon

- le type du variable dépendante
 - variable quantitative – régression linéaire
 - variable dichotomique – régression logistique
- la linéarité de la fonction
 - régression linéaire
 - régression non linéaire
- le nombre de variables dépendantes:
 - régression univariée (une variable dépendante)
 - régression multivariée (≥ 2 variables dépendantes)
- le nombre de variables indépendantes:
 - régression **simple** (une variable indépendante)
 - régression **multiple** (≥ 2 variables indépendantes)

23

23

Régression linéaire simple

But étudier la relation linéaire entre deux variables, depuis lesquels la variable dépendante est quantitative

Objectifs:

- **prédire les valeurs** d' une variable quantitative en fonction des valeurs de l'autre variable
- **évaluer**
 - **l'existence** d'une association/lien/relation entre les deux variables
 - **la direction** de l' association/lien/relation entre deux variables
 - **l' importance** de l' association/lien/relation entre deux variables

24

24

Régression linéaire simple

Conditions d'application:

- variable dépendante quantitative
- observations indépendants
- existence d'une relation linéaire entre les deux variables (X et Y)
- erreurs normalement distribués (résidus)
- homoscédasticité - variance constante des erreurs (résidus)

25

25

Régression linéaire simple - calcul

$Y = b_0 + b_1 X + \varepsilon$ ou b_0, b_1 = coefficients de la régression

- Méthode de calcul des coefficients b_0, b_1
= méthode des moindres carrés
- critère de la méthode: minimisé la somme de tous les carrés de la distance du chaque point par rapport à la droite:

$$\sum_{i=1}^n (Y_i^{predite} - Y_i)^2 Y_i^{predite} = b_0 + b_1 X_i$$

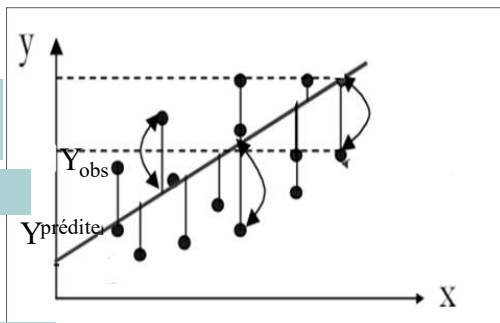
-on obtient:

$$b_1 = \frac{COV(X,Y)}{S_X^2}$$

Valeur observée
covariance

$$b_0 = \bar{Y} - a_1 \bar{X}$$

Moyenne du Y, moyenne du X



26

26

Régression linéaire simple

- **Interprétation**

- La droite de régression $Y(X)$: $Y = b_0 + b_1 X$

b_0 = est l'ordonnée à l'origine – la valeur du Y quand X est égal à 0 (d'habitude cette information n'est pas utile pour les médecins, elle présente une situation qui en réalité est impossible)

b_1 = la pente de la droite de régression.

Interprétation de b_1 - du coefficient de la variable X

chaque unité de mesure de la variable indépendante - X en plus augmente en moyenne la variable dépendante - Y avec la valeur du coefficient de la variable indépendante X – b_1

27

27

Régression linéaire simple

$$Y = b_0 + b_1 X$$

Interprétation de b_1 - du coefficient de la variable X

- $b_1 > 0$ tendance croissante/ pente ascendante/ pente positive/ lien direct proportionnel
- $b_1 < 0$ tendance décroissante/ pente descendante/ pente négative/ lien inversement proportionnel
- $b_1 \cong 0 \Rightarrow$ aucune tendance
- plus b_1 est grand (en valeur absolue) plus la relation est forte
- plus b_1 est proche du 0, plus la relation est faible

28

28

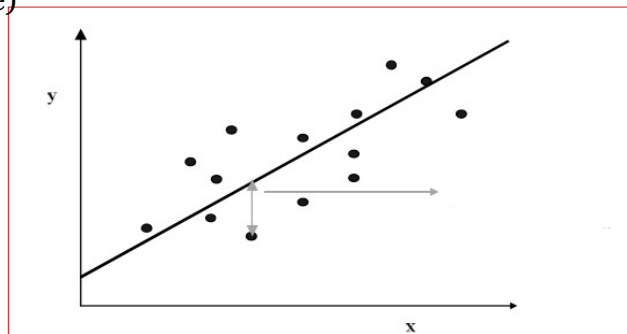
Test statistique de signifiante pour le coefficient de la pente (b_1) de la régression linéaire

29

- L'hypothèse nulle:
 - il n'y a pas une différence statistiquement significative entre le coefficient (la pente) de la variable x et 0
 - le coefficient (la pente) de la variable x n'est pas statistiquement significative
- L'hypothèse alternative:
 - il y a une différence statistiquement significative entre le coefficient (la pente) de la variable x et 0
 - le coefficient (la pente) de la variable x est statistiquement significative
- La décision avec p-value.
 - Si $p\text{-value} < \alpha (=0,05) \Rightarrow$ on rejette H_0 et on accepte H_1
 - Si $p\text{-value} \geq \alpha (=0,05) \Rightarrow$ on ne peut pas rejeter H_0

29

Résidus = Ecart entre chaque valeur Y_i observée et chaque valeur prédite par la droite (écart non expliqué par la droite)



Variance Résiduelle (des résidus) S_R^2 = une mesure de la dispersion des points autour de la droite de régression:

$$S_R^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - Y_i^{\text{prédite}})^2$$

30

30

Le coefficient de détermination

Variance totale de la variable dépendante =

- Variance donnée par la droite (variance de régression) + variance NON donnée par la droite (variance résiduelle)

$$r^2 = \frac{\text{Variance}_{\text{régression}}}{\text{Variance}_{\text{totale}}}$$

- Nommée : **coefficient de détermination**
- Notation: **d** = **r²** (le carré du coefficient de corrélation Pearson)
- nombre de 0 à 1. On le transforme dans pourcentage pour l'interpréter (0 → 100%)

Interprétation: pourcentage de la variance (variation) d'une variable dépendante explique par la relation linéaire avec la variable indépendante

- plus d est grand plus la relation/association/liens est forte
- plus d est proche de 0, plus la relation est faible

31

31

Le coefficient de détermination

Exemple :

On a fait une régression entre X = Poids (kg) et Y = HDL Cholestérol (mg/dl). La relation entre les deux est linéaire, les variables sont normalement distribuées, les observations sont indépendantes, les erreurs normalement distribués, et homoscédastiques. On a trouvé le coefficient de détermination = 0,82.

Interprétez le coefficient de détermination

d = 0,82 ⇒ **82% de la variance (variation) du cholestérol 'peut être expliquée par la relation linéaire avec le Poids.**
Le reste de 18% est due à des autres facteurs.

32

32

Corrélation, régression, et causalité

Le fait qu'on trouve dans une étude une corrélation statistiquement significative, ou un coefficient de détermination importante, ou si le coefficient de la régression linéaire est statistiquement significative, ne signifie pas fortement que la relation est causale !!!

Ces informations, nous indique qu'il y a des relations, entre les variables, mais on ne sait pas qui est la cause et qui est l'effet.

Pour montrer la causalité, la méthodologie de la réalisation de l'étude est très importante. On va voir ça dans l'année prochaine.

33

Régression linéaire multiple

But étudier la relation linéaire entre plus de deux variables et une variable dépendante quantitative

Objectifs:

- **prédire les valeurs** d'une variable quantitative en fonction des valeurs des **plusieurs** autres **variables**
- **évaluer**
 - **l'existence** d'une association/lien/relation entre les deux variables
 - **la direction** de l' association/lien/relation entre deux variables
 - **l' importance** de l' association/lien/relation entre deux variables
- **ajuster/corriger/tenir compte** - l'effet d' autre variables

34

34

Régression linéaire multiple - Quantification de l'importance de la relation pour plusieurs variables

Les **relations entre différents variables médicales** sont souvent **complexes**, et plusieurs facteurs/ variables/ caractéristiques sont impliquées.

Avec le **coefficient de corrélation** ou avec la **régression linéaire simple** on peut **évaluer** la relation entre **seulement deux variables** – donc une analyse uni varie (entre une variable dépendante – et une variable explicative/ indépendante)

Mais on a **besoin de tenir compte de l'effet d'autres variables** (nommée de confusion, ou autres variables explicatives) qui sont connues qu'ils influence la relation qu'on étudie. On doit faire une étude bibliographique pour identifier ces variables de confusion.

Pour **tenir compte des plusieurs variables** on doit utiliser des **techniques multi variées** (eg. **Régression linéaire multiple**,³⁵ régression logistique multiple,)

35

Régression linéaire multiple - Quantification de l'importance de la relation pour plusieurs variables

S'il y a plusieurs variables qui sont liée a la variable dépendante d'intérêt, on peut évaluer leur importance avec une **régression linéaire multiple** qui offre une coefficient (nomme ajusté – « adjusted » en anglais) pour chaque variable indépendante:

La variable **dépendante** (predite,expliquee) de la régression:

Les triglycérides (mg/dL)

Les variables **indépendantes** (explicatives,predictives,explicatives)

Qualitatives (facteurs) (ex. le sexe)

Quantitatives (ex. le poids (kg))

Le **coefficient ajusté** peut nous **rapprocher plus a la vérité que le coefficient brut** (« crude »/ » unadjusted» en anglais - sans ajustement) d'un régression linéaire simple, ou une coefficient de corrélation simple, parce que **on peut tenir compte d'autres variables** qui agit dans le même temps ³⁶

36

Régression linéaire multiple - Quantification de l'importance de la relation pour plusieurs variables

L' équation du **régression linéaire multiple**:

Variable dépendante= coefficient_1 * variable_1 + coefficient_2 * variable_2 + ...
+ coefficient_n * variable_n + coefficient_0

Ex: triglycérides (mg/dL) = 23,10 * obésité (oui/non) + 1,14 * cholestérol (mg/dL)

L' interprétation du coefficient ajusté (adjusted – en anglais) pour des variables **Qualitatives dichotomiques** (ex. obésité):

l'augmentation de la variable dépendante – les triglycérides en moyenne (ici il est de 23,1 mg/dL) pour ceux qui ont le facteur présent (être obèse - la variable indépendante) comparée a ceux qui n'ont pas le facteur (ne sont pas obèses), si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – le cholestérol)

ceux qui ont le facteur présent (être obèse - la variable indépendante) ont la variable dépendante – les triglycérides en moyenne plus grand avec 23,1 mg/dL comparée a ceux qui n'ont pas le facteur (ne sont pas obèses), si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – le cholestérol)

37

Régression linéaire multiple - Quantification de l'importance de la relation pour plusieurs variables

L' équation du **régression linéaire multiple**:

Variable dépendante= coefficient_1 * variable_1 + coefficient_2 * variable_2 + ...
+ coefficient_n * variable_n + coefficient_0

Ex: triglycérides (mg/dL) = 23,10 * obésité (oui/non) + 1,14 * cholestérol (mg/dL)

L' interprétation du coefficient ajusté (adjusted – en anglais) pour des variables **Quantitatives** (ex. le cholestérol):

l'augmentation de la variable dépendante – les triglycérides en moyenne (ici il est de 1,14 mg/dL) pour chaque changement d'un unité de mesure de la variable indépendante (chaque 1 mg/dL du cholestérol) si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – l' obésité)

pour chaque unité de mesure de la variable indépendante (chaque 1 mg/dL du cholestérol) en plus, la variable dépendante – les triglycérides en moyenne augmente avec 1,14 mg/dL si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – l' obésité)

38

Test statistique de signifiante pour le coefficient de $(b_1, 2, \dots, n)$ chaque variable indépendante de la régression linéaire multiple

39

- L'hypothèse nulle:
 - il n'y a pas une différence statistiquement significative entre le coefficient de la pente du variable $x_1 \dots x_n$ et 0
 - le coefficient de la variable $x_1 \dots x_n$ du régression n'est pas statistiquement significative
- L'hypothèse alternative:
 - il y a une différence statistiquement significative entre le coefficient de la pente du variable $x_1 \dots x_n$ et 0
 - le coefficient de la variable $x_1 \dots x_n$ du régression est statistiquement significative
- La décision avec p-value.
 - Si $p\text{-value} < \alpha (=0,05) \Rightarrow$ on rejete H_0 et on accepte H_1
 - Si $p\text{-value} \geq \alpha (=0,05) \Rightarrow$ on ne peut pas rejeter H_0

39

Logiciel pour l'analyse de corrélations et de régressions

- EPIINFO (voir le tp)
 - Analysis, commandes
 - REGRESSION
 - Scatter
- EXCEL(voir le tp)
 - Fonction CORREL
 - Chart, scatter, trendline
 - Menu Tools, commande Data Analysis
 - Regression
- R (R Commander)
 - Statistics / Summaries / Correlation coefficient/test
 - Statistics / Models / Linear model

40

40

Récapitulatif

Type des variables	Nature des données	Coefficient de corrélation	Formule du coefficient
quantitative	normale distribuées	Pearson (r)	$COV(X,Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ $r = \frac{COV(X,Y)}{S_X \cdot S_Y}$
quantitative	non normale distribuées	Spearman (ρ - rho)	$\rho = r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \text{ ou } d_i = R_{x_i} - R_{y_i}$
qualitative ordinales	-	Spearman (ρ - rho)	$\rho = r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \text{ ou } d_i = R_{x_i} - R_{y_i}$

X_i, Y_i – sont les valeurs des deux séries des données. \bar{X} et \bar{Y} sont les moyennes des deux séries. R_{x_i} et R_{y_i} sont les rangs des valeurs X_i et Y_i après leur rangement dans ordre croissant. n – nombre des observations. S_X et S_Y sont les déviations standard d'échantillonnage. $COV(X,Y)$ – la covariance

41

Comment identifier la distribution des données en regardant un graphique nouage des points, et d'identifier le bon coefficient de corrélation à utiliser

Si on regarde une **graphique** de type **nouage des points**, on peut **deviner approximative** si la **distribution** des deux variables (si c'est une distribution **normale** (gaussienne) ou pas). On sait que la **distribution normale** est **symétrique** autour de la moyenne, et on peut observer ça dans le graphique.

Ex. **Si les deux variables sont normale distribuées** le **graphique** a **approximative** la forme d'une **ellipse**.

42

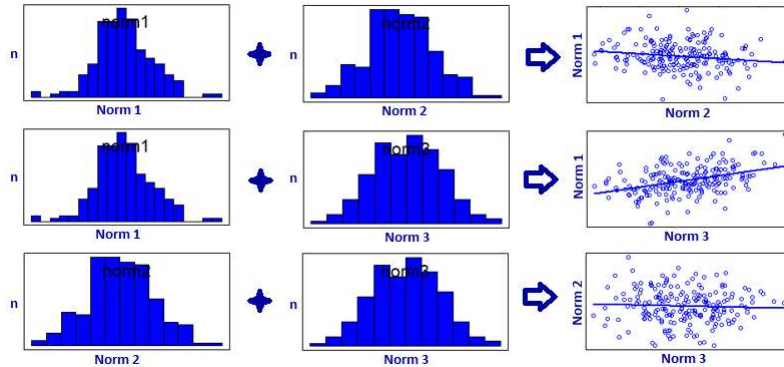
Comment identifier la distribution des données en regardant un graphique nuage des points, et d'identifier le bon coefficient de corrélation à utiliser

Ex. Ici il y a **3 variables**: Norm 1, Norm 2, et Norm 3 qui sont **normale distribuées** (come on peut observer sur les histogrammes).

Pour chaque combinaison entre deux variables Norm 1 + 2, Norm 1 + 3, et Norm 2 + 3 il y a une nuage des points.

On peut observer que **si les deux variables sont normale distribuées** le **graphique a la forme d'une ellipse** (n'importe si la tendance est décroissante, croissante, ou sans).

Donc **si vous observez des nuages des points en forme d'ellipse, et la relation est approximative lineaire** vous pouvez deviner que c'est une **suggestion** que les **deux variables** sont **normale distribuées**, et on peut utiliser le **coefficient de corrélation Pearson**.



43

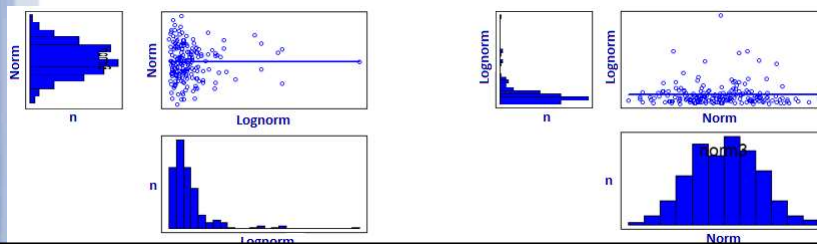
Comment identifier la distribution des données en regardant un graphique nuage des points, et d'identifier le bon coefficient de corrélation à utiliser

Ex. Ici il y a **2 variables**: la variable **Norm** qui est **normale distribuée** (come on peut observer sur l'histogramme – symétrique, forme gaussienne – de chapeau), et la variable **Lognorm** – qui **n'est pas normale distribuée**, ayant une forte **asymétrie à droite** (come on peut observer sur l'histogramme – n'est pas symétrique, il y a une **queue à la droite**).

On peut observer deux nuages des points: sur axe X – Lognorm, sur axe Y – Norm sur axe X – Norm, sur axe Y – Lognorm

On peut observer que **si une des deux variables ne sont pas normale distribuées** le **graphique n'a pas la forme d'une ellipse**. On observe a dans le nuage des points a gauche que pour la **variable Norm** les données sont **approximative symétrique** distribuées autour de centre du nuage sur l'axe verticale, mais la **variable Lognorm** a la plupart des points a gauche, et de moins en moins des points, de plus en plus distancées vers la droite du nuage des points, donc une **distribution asymétrique (positive, avec une queue a la droite)**. Le graphique nuage des points a la droite est identique, mais avec une rotation, parce que les mêmes variables sont inversées sur les axes.

Donc **si vous observez des nuages des points qui n'est pas en forme d'ellipse** vous pouvez deviner que c'est une **suggestion** que au moins une des deux **variables ne sont pas normale distribuées**, et on c'est une suggestion qu'on **ne peut pas utiliser** le **coefficient de corrélation Pearson**, on doit utiliser plutôt le **coefficient de corrélation Spearman**.



44

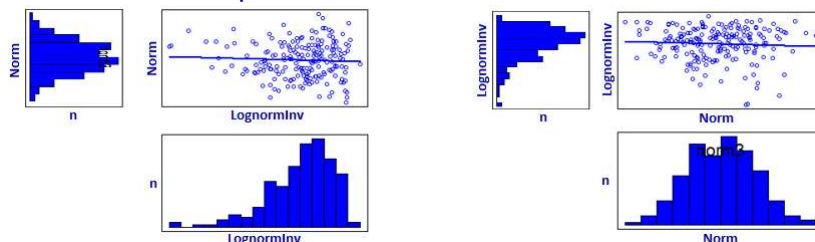
Comment identifier la distribution des données en regardant un graphique nouage des points, et d'identifier le bon coefficient de corrélation à utiliser

Ex. Ici il y a **2 variables Norm** qui est **normale distribuée** (come on peut observer sur l'histogramme – symétrique, forme gaussienne – de chapeau), et la variable **LognormInv** – qui **n'est pas normale distribuée**, ayant une forte **asymétrie a gauche** (come on peut observer sur l'histogramme – n'est pas symétrique, il y a une **queue a gauche**).

On peut observer deux nouages des points: sur axe X – LognormInv, sur axe Y – Norm
sur axe X – Norm, sur axe Y – LognormInv

On peut observer que si **une des deux variables ne sont pas normale distribuées** le **graphique n'a pas la forme d'une ellipse**. On observe a dans le nouage des points a gauche que pour la **variable Norm** les données sont **approximative symétrique** distribuées autour de centre du nuage sur l'axe verticale, mais la **variable LognormInv** a la plupart des points a droite, et de moins en moins des points, de plus en plus distancées vers la gauche du nuage des points, donc une **distribution asymétrique (négative, avec une queue a la gauche)**. Le graphique nuage des points a la droite est identique, mais avec une rotation, parce que les mêmes variables sont inversées sur les axes.

Donc **si vous observez des nouages des points qui n'est pas en forme d'ellipse** vous pouvez deviner que c'est une **suggestion** que au moins une des deux **variables ne sont pas normale distribuées**, et on c'est une suggestion qu'on **ne peut pas utiliser** le **coefficient de corrélation Pearson**, on doit utiliser plutôt le **coefficient de corrélation Spearman**.



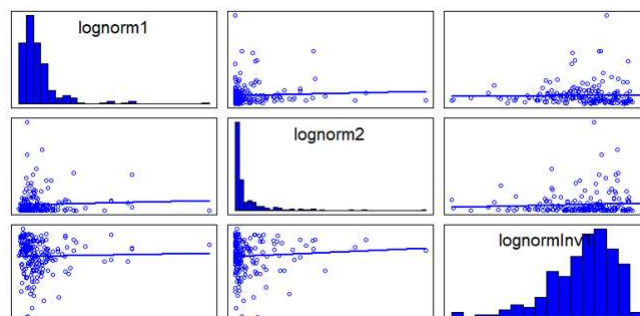
45

Comment identifier la distribution des données en regardant un graphique nouage des points, et d'identifier le bon coefficient de corrélation à utiliser

Ex. Ici il y a **3 variables**: les variables **Lognorm 1, et 2** – qui **ne sont pas normale distribuées**, ayant une forte **asymétrie a droite** (come on peut observer sur l'histogramme – n'est pas symétrique, il y a une **queue a la droite, positive**), et la variable **LognormInv** – qui **n'est pas normale distribuée**, ayant une forte **asymétrie a gauche, negative** (come on peut observer sur l'histogramme – n'est pas symétrique, il y a une **queue a gauche**).

On peut observer **une matrice des nouages des points**, et sur la **diagonale** sont les **histogrammes des variables**.

On peut observer que si **les deux variables ne sont pas normale distribuées** le **graphique n'a pas la forme d'une ellipse**. Donc **si vous observez des nouages des points qui n'est pas en forme d'ellipse** vous pouvez deviner que c'est une **suggestion** que au moins une des deux **variables ne sont pas normale distribuées**, et on c'est une suggestion qu'on **ne peut pas utiliser** le **coefficient de corrélation Pearson**, on doit utiliser plutôt le **coefficient de corrélation Spearman**.

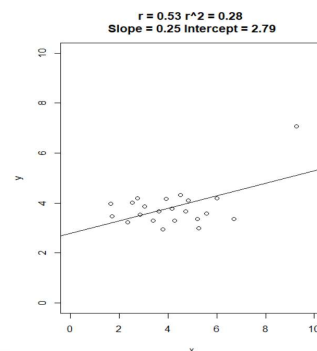
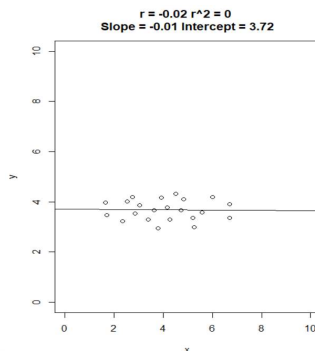


46

L'influence des valeurs aberrantes sur le coefficient de corrélation Pearson, et sur la droite de régression

Ex. Dans le **graphique a gauche** on peut observer une nouage des points en forme d'ellipse, avec une relation linéaire. Pour ce graphique le **coefficient de corrélation Pearson** est **-0,02**, et la **pente** de la droite de régression est **-0,01**, et l'intercept **3,72**. Donc on peut observer qu'il **n'y a pas de relation** apparente entre les variables.

Dans le **graphique a droite** on peut observer une le même nouage des points, mais avec une **valeur aberrante a la droite et en haut**. Pour ce graphique le **coefficient de corrélation Pearson** est **0,53**, et la **pente** de la droite de régression est **0,25**, et l'intercept **2,79**. Si on regarde seulement le coefficient et la pente on peut penser que il y a une corrélation, une relation entre les deux variables, mais cette relation est induite d'une manière fausse par la valeur aberrante (qui est aussi une point de levier). S'il vous plait de **ne pas faire confiance dans les statistiques dans ce gens de situations**. Les auteurs de l'article doivent évaluer la cause de ces valeurs, et aussi l'effet de la ou les valeurs aberrantes. Au lieu du coefficient de corrélation Pearson, ici sera **mieux d'utiliser le coefficient de corrélation Spearman**.

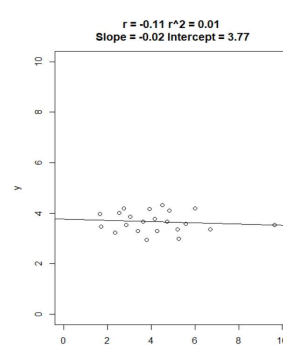
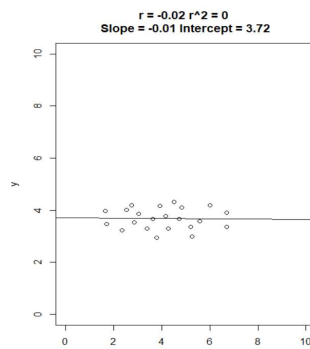


47

L'influence des valeurs aberrantes sur le coefficient de corrélation Pearson, et sur la droite de régression

Ex. Dans le **graphique a gauche** on peut observer une nouage des points en forme d'ellipse, avec une relation linéaire. Pour ce graphique le **coefficient de corrélation Pearson** est **-0,02**, et la **pente** de la droite de régression est **-0,01**, et l'intercept **3,72**. Donc on peut observer qu'il **n'y a pas de relation** apparente entre les variables.

Dans le **graphique a droite** on peut observer une le même nouage des points, mais **avec une valeur aberrante a la droite**. Pour ce graphique le **coefficient de corrélation Pearson** est **-0,11**, et la **pente** de la droite de régression est **-0,02**, et l'intercept **3,77**. Ici la valeur aberrante n'a pas influencée beaucoup les statistiques (elle n'a pas une effet de levier). S'il vous plait de **ne pas faire confiance dans les statistiques dans ce gens de situations**. Les auteurs de l'article doivent évaluer la cause de ces valeurs, et aussi l'effet de la ou les valeurs aberrantes. Au lieu du coefficient de corrélation Pearson, ici sera **mieux d'utiliser le coefficient de corrélation Spearman**.

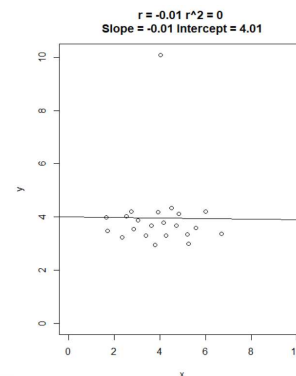
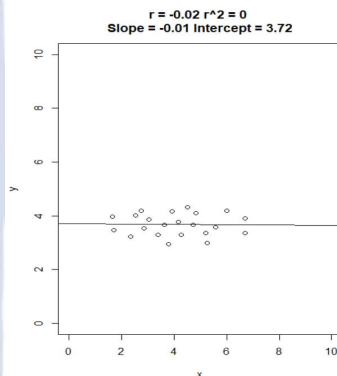


48

L'influence des valeurs aberrantes sur le coefficient de corrélation Pearson, et sur la droite de régression

Ex. Dans le **graphique à gauche** on peut observer un nuage des points en forme d'ellipse, avec une relation linéaire. Pour ce graphique le **coefficient de corrélation Pearson** est **-0,02**, et la **pente** de la droite de régression est **-0,01**, et le **coefficient libre** est **3,72**. Donc on peut observer qu'il n'y a **pas de relation** apparente entre les variables.

Dans le **graphique à droite** on peut observer le même nuage des points, mais avec une **valeur aberrante au milieu, et en haut**. Pour ce graphique le **coefficient de corrélation Pearson** est **-0,01**, et la **pente** de la droite de régression est **-0,01**, et le **coefficient libre** est **4,01**. Ici la valeur aberrante n'a pas influencé beaucoup les statistiques (elle n'a pas un effet de levier), sauf le coefficient libre. S'il vous plaît de **ne pas faire confiance dans les statistiques dans ce genre de situations**. Les auteurs de l'article doivent évaluer la cause de ces valeurs, et aussi l'effet de la ou les valeurs aberrantes. Au lieu du coefficient de corrélation Pearson, ici sera **mieux d'utiliser le coefficient de corrélation Spearman**.



49

Exemples d'articles scientifiques avec tests statistiques

Coefficient de corrélation Pearson, régression linéaire simple

Student Attendance and Academic Performance in Undergraduate Obstetrics/Gynecology Clinical Rotations FREE

Richard P. Deane, MB BCH¹; Deirdre J. Murphy, MD¹ JAMA. 2013;310(21):2282-2288. doi:10.1001/jama.2013.282228.

<http://journalgateway.com/ijomp/article/view/583>

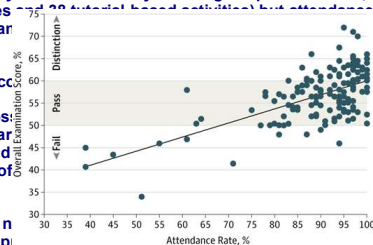
Objective To evaluate the relationship between student attendance and academic performance in a medical student obstetrics/gynecology clinical rotation.

Design, Setting, and Participants A prospective cohort study of student attendance at clinical and tutorial-based activities during a full academic year (September 2011 to June 2012) within a publicly funded university teaching hospital in Dublin, Ireland. Students were expected to attend 64 activities (26 clinical activities and 38 tutorial-based activities) but attendance was not mandatory. All 147 fourth-year medical students who completed an examination were included.

Exposures Student attendance at clinical and tutorial-based activities, recorded as a percentage.

Main Outcomes and Measures The overall examination score (out of a possible 100 points), an end-of-year choice questions (40 points) and 6 short-answer questions (40 points), and examination (80 points). Students were required to have an overall score of the long-case clinical/oral examination (50%) to pass.

Results The mean attendance rate was 89% (range, 39%-100% [SD, 11%], n = 147) and students who failed an end-of-year examination performed significantly lower rates. There was a positive correlation between attendance and overall examination score ($r = 0.59$ [95% CI, 0.44-0.70]; $P < .001$). Both clinical attendance ($r = 0.50$ [95% CI, 0.32-0.64]; $P < .001$) and tutorial-based attendance ($r = 0.57$ [95% CI, 0.40-0.70]; $P < .001$) were positively correlated with overall examination score. The associations persisted after controlling for confounding factors of student sex, age, country of origin, previous failure in an end-of-year examination, and the timing of the rotation during the academic year. Distinction grades (overall score of $\geq 60\%$) were present only among students with attendance rates of 80% or higher. The odds of a distinction grade increased with each 10% increase in attendance (adjusted odds ratio, 5.52; 95% CI, 2.17-14.00). The majority of failure grades (6/10 students; 60%) occurred in students with attendance rates lower than 80%. The adjusted odds ratio for failure with attendance rates of 80% or higher was 0.11 (95% CI, 0.02-0.72).



50

Exemples des questions pour l'examen

	B unadjusted	95% CI	P-value	B adjusted	95% CI	P-value
Homme	22,3	(18,3-26,3)	<0,001	18,7	(14,6-22,8)	<0,001
Âge (annees)	1,14	(1,07-1,21)	<0,001	1,08	(1,02-1,14)	<0,001

1) Quelles affirmations, en concernant la régression linéaire simple (uni variée) et multiple (multivariée) dans le tableau pour prédire la circonférence abdominale (cm) en fonction de sexe et l'âge, sont correctes:

- le coefficient ajustée pour la variable sexe est 18,7
- le coefficient ajustée pour la variable âge est 1,08
- le coefficient crude/brut pour la variable sexe est 22,3
- le coefficient unadjusted pour la variable âge est 1,14
- un homme a en moyenne 22,3 cm en plus de circonférence abdominale par rapport a une femme avec ajustement pour la variable âge
- chaque unité de mesure de la variable âge (chaque année) en plus augmente la circonférence abdominale en moyenne avec 1,14 cm sans ajustement pour la variable sexe
- chaque unité de mesure de la variable âge (chaque année) en plus augmente la circonférence abdominale en moyenne avec 1,08 cm avec ajustement pour la variable sexe
- chaque unité de mesure de la variable âge (chaque année) en plus augmente la circonférence abdominale en moyenne avec 1,08 cm en contrôlant pour la variable sexe
- chaque unité de mesure de la variable âge (chaque année) en plus augmente la circonférence abdominale en moyenne avec 1,08 cm en tenant constante la variable sexe
- un homme a en moyenne 18,7 cm en plus de circonférence abdominale par rapport a une femme sans contrôler pour la variable âge
- le coefficient de la variable sexe du régression simple est statistiquement significative
- le coefficient de la variable sexe du régression multiple n'est pas statistiquement significative
- le coefficient de la variable âge du régression simple est statistiquement significative

Réponse: a, b, c, d, e, f, g, h, i, j, k, m

51

Exemples des questions pour l'examen

	B unadjusted	95% CI	P-value	B adjusted	95% CI	P-value
Homme	22,3	(18,3-26,3)	<0,001	18,7	(14,6-22,8)	<0,001
Âge (annees)	1,14	(1,07-1,21)	<0,001	1,08	(1,02-1,14)	<0,001

1) Quelles affirmations, en concernant la régression linéaire simple (uni variée) et multiple (multivariée) dans le tableau pour prédire la circonférence abdominale (cm) en fonction de sexe et l'âge, sont correctes:

Explications pour l'exemple:

Le tableau montre trois régressions linéaires:

- Deux régressions linéaires simples - uni variée (parce que il y a la colonne B unadjusted (le coefficient, la pente de la régression, non ajuste/ crue – non adjusted / crude – en anglais), avec deux variables homme et âge; et une régression linéaire multiple – multi variée (parce que il y a la colonne B adjusted (le coefficient, la pente de la régression, ajuste – adjusted – en anglais)
- On a une régression linéaire simple (uni variée) qui prédit la circonférence abdominale (cm) en fonction du sexe – une variable qualitative dichotomique (l' équation de la régression est: circonférence abdominale (cm) = 22,3 * home + coefficient libre (non montrée dans le tableau)). Le coefficient non ajuste (unadjusted en anglais) du sexe est 22,3. Donc ceux qui ont le facteur présent (être home - la variable indépendante) ont la variable dépendante – la circonférence abdominale en moyenne plus grand avec 22,3 cm comparée a ceux qui n'ont pas le facteur (les femmes). La valeur du p du coefficient de la variable home est <0,001 – donc le coefficient de la variable sexe du régression est statistiquement significative – il y a une différence statistiquement significative entre le coefficient de la pente du variable sexe et 0. La colonne 95% CI montre l'intervalle de confiance 95% du coefficient. Par exemple pour le coefficient non ajustée des hommes – 22,3, la vrai valeur du coefficient dans la population se trouve entre 18,3 et 26,3 avec une probabilité de 95%.

52

Exemples des questions pour l'examen

	B unadjusted	95% CI	P-value	B adjusted	95% CI	P-value
Homme	22,3	(18,3-26,3)	<0,001	18,7	(14,6-22,8)	<0,001
Âge (années)	1,14	(1,07-1,21)	<0,001	1,08	(1,02-1,14)	<0,001

- 1) Quelles affirmations, en concernant la régression linéaire simple (uni variée) et multiple (multivariée) dans le tableau pour prédire la circonférence abdominale (cm) en fonction de sexe et l'âge, sont correctes:

Explications pour l'exemple:

- On a une régression linéaire simple (uni variée) qui prédit la circonférence abdominale (cm) en fonction de l'âge – une variable quantitative (l'équation de la régression est: circonférence abdominale (cm) = 1,14 * âge + coefficient libre (non montrée dans le tableau)). Le coefficient non ajusté (unadjusted en anglais) de l'âge est 1,14. Donc pour chaque unité de mesure de la variable indépendante (chaque 1 année de l'âge) en plus, la variable dépendante – la circonférence abdominale en moyenne augmente avec 1,14 cm. La valeur du p du coefficient de la variable âge est <0,001 – donc le coefficient de la variable âge du régression est statistiquement significative – il y a une différence statistiquement significative entre le coefficient de la pente du variable âge et 0.

53

Exemples des questions pour l'examen

	B unadjusted	95% CI	P-value	B adjusted	95% CI	P-value
Homme	22,3	(18,3-26,3)	<0,001	18,7	(14,6-22,8)	<0,001
Âge (années)	1,14	(1,07-1,21)	<0,001	1,08	(1,02-1,14)	<0,001

- 1) Quelles affirmations, en concernant la régression linéaire simple (uni variée) et multiple (multivariée) dans le tableau pour prédire la circonférence abdominale (cm) en fonction de sexe et l'âge, sont correctes:

Explications pour l'exemple:

- On a une régression linéaire multiple (multi variée) qui prédit la circonférence abdominale (cm) en fonction de sexe et de l'âge – une variable qualitative dichotomique et une variable quantitative (l'équation de la régression est: circonférence abdominale (cm) = 18,7 * home + 1,08 * âge (années) + coefficient libre (non montrée dans le tableau)).
 - Le coefficient ajuste (adjusted en anglais) de l'âge est 1,08. Donc pour chaque unité de mesure de la variable indépendante (chaque 1 année de l'âge) en plus, la variable dépendante – la circonférence abdominale en moyenne augmente avec 1,08 cm si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – le sexe). La valeur du p du coefficient de la variable âge est <0,001 – donc le coefficient de la variable âge du régression est statistiquement significative – il y a une différence statistiquement significative entre le coefficient de la pente du variable âge et 0.
 - Le coefficient ajuste (adjusted en anglais) du sexe est 18,7. Donc ceux qui ont le facteur présent (être home - la variable indépendante) ont la variable dépendante – la circonférence abdominale en moyenne plus grand avec 18,7 cm comparée à ceux qui n'ont pas le facteur (les femmes) si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – l'âge). La valeur du p du coefficient de la variable home est <0,001 – donc le coefficient de la variable sexe du régression est statistiquement significative – il y a une différence statistiquement significative entre le coefficient de la pente du variable sexe et 0.

54

Exemples des questions pour l'examen

2) L'équation de régression de la tension artérielle systolique (mmHg) = la épaisseur de la paroi ventriculaire (mm) * 2,8 + 84,5. Le coefficient de corrélation Pearson = 0,63. Lesquelles des réponses suivantes sont correctes :

- a) la covariance est négatif
- b) pour chaque millimètre en plus de épaisseur de la paroi ventriculaire, la tension artérielle systolique augmente en moyenne avec 2,8 mmHg
- c) la relation entre les deux variables est inversement proportionnelle.
- d) la corrélation est moyenne vers bonne
- e) 84,5 mmHg est l'ordonnée à l'origine

Réponse: b, d, e

Explications:

- a) La covariance a le même signe avec le coefficient de corrélation => positive
- b) L'interprétation classique du coefficient d'une variable quantitative
- c) Le signe du coefficient de corrélation Pearson (0,64=>positive), le signe de la pente (2,8=>positive), donc il y a une relation directe proportionnelle.
- d) Oui, parce que le coefficient de corrélation Pearson = 0,63, se trouve entre 0,5 et 0,75, l'intervalle précisée par Colton, et donc la corrélation est moyenne vers bonne
- e) Le coefficient libre est l'ordonnée à l'origine

DL1

55

Exemples des questions pour l'examen

3) L'équation de régression de la tension artérielle systolique (mmHg) = la épaisseur de la paroi ventriculaire (mm) * 2,7 + 79,2. Le coefficient de détermination = 0,36 Lesquelles des réponses suivantes sont correctes :

- a) le coefficient de corrélation est < 0
- b) pour chaque millimètre en plus de épaisseur de la paroi ventriculaire, la tension artérielle systolique augmente en moyenne avec 2,7 mmHg
- c) la relation entre les deux variables est proportionnelle
- d) la corrélation est moyenne vers bonne
- e) 36 % de la variation de la tension artérielle systolique peut être expliquée par la relation linéaire avec la épaisseur de la paroi ventriculaire

Réponse: b, c, d, e

Explications:

- a) le coefficient de corrélation Pearson a le même signe avec le coefficient de la variable indépendante (2,7) => positive
- b) L'interprétation classique du coefficient d'une variable quantitative
- c) Le signe de la pente (2,8=>positive), donc il y a une relation directe proportionnelle.
- d) Oui, parce que le coefficient de corrélation Pearson, est la racine carrée du coefficient de détermination (qui est le carré du coefficient de corrélation Pearson), donc = 0,60, se trouve entre 0,5 et 0,75, l'intervalle précisée par Colton, et donc la corrélation est moyenne vers bonne
- e) L'interprétation classique du coefficient de détermination

56

Exemples des questions pour l'examen

4) Le poids et le cholestérol des 60 malades a été observée. Les suivantes statistiques ont été calculées. Pour le poids: coefficient d'asymétrie=0,28, coefficient d'aplatissement=0,7, coefficient de variation = 0,40. Pour le cholestérol: coefficient d'asymétrie=2,2, coefficient d'aplatissement=0,5. Le coefficient de corrélation Pearson entre le poids et le cholestérol est: 0,30. La valeur p pour le coefficient de corrélation et la régression est <0,01. Lesquelles des réponses suivantes sont correctes :

- a) le poids est relativement hétérogène
- b) on doit calculer le coefficient de corrélation Spearman
- c) le coefficient de détermination est 0,09
- d) la distribution du cholestérol a une queue a la gauche
- e) 40 % de la variance (variation) du poids peut être expliquée par la relation linéaire avec le cholestérol

Réponse: b, c

Explications:

- a) Le niveau d' hétérogénéité d'une variable est identifiable a l'aide du coefficient de variation (écart type / moyenne) – voir le cours avec les indicateurs de dispersion. 0,40 représente 40%, qui est plus grand que 30%, donc les données sont hétérogènes (un intervalle entre 20% et 30% indique une série relativement hétérogène) => réponse a) est faux
- b) Le coefficient de corrélation Pearson peut être utilise pour deux variables quantitatives normale distribuées. Ici, le poids est normale distribuée, mais le cholestérol n'est pas normale distribuée. On considère des données normale distribuées, quand le coefficient d' asymétrie et le coefficient d' aplatissement se trouvent entre [-1, 1] – voir le cours avec l' évaluation de la normalité des données. Le coefficient de corrélation Spearman peut être utilise pour: 2 variables quantitatives non normale distribuées; pour une variable normale distribuée, et un autre non normale distribuée; pour deux variables ordinales; pour une variable ordinale et une variable quantitative (n'importe s'il est normale distribuée, ou non). Donc Spearman est notre choix ici
- c) Le coefficient de détermination est le carré du coefficient de corrélation Pearson = $0,3 * 0,3 = 0,09$
- d) Non, parce que le coefficient d' asymétrie est >0 (2,2), qui indique une queue a la droite
- e) L' interprétation classique du coefficient de détermination, mais au lieu de 40% la bonne valeur est 9% (0,09 – voir en haut)

57

Exemples des questions pour l'examen

5) Le poids et le HDL cholestérol des 60 malades a été observée. Les suivantes statistiques ont été calculées. Pour le poids: coefficient d'asymétrie=0,28, coefficient d'aplatissement=0,7, coefficient de variation = 0,40. Pour le HDL cholestérol: coefficient d'asymétrie= -0,2, coefficient d'aplatissement=0,5. Le coefficient de corrélation Pearson entre le poids et le HDL cholestérol est: -0,30. La valeur p pour le coefficient de corrélation et la régression est <0,01. Lesquelles des réponses suivantes sont correctes :

- a) le poids est hétérogène
- b) on doit calculer le coefficient de corrélation Spearman
- c) le coefficient de détermination est 0,09
- d) la distribution du HDL cholestérol a une queue a la gauche
- e) 9 % de la variance (variation) du poids peut être expliquée par la relation linéaire avec le HDL cholestérol

Réponse: a, c, d, e

Explications:

- a) Le niveau d' hétérogénéité d'une variable est identifiable a l'aide du coefficient de variation (écart type / moyenne) – voir le cours avec les indicateurs de dispersion. 0,40 représente 40%, qui est plus grand que 30%, donc les données sont hétérogènes => réponse a) est correcte
- b) Le coefficient de corrélation Pearson peut être utilise pour deux variables quantitatives normale distribuées. Ici, le poids est normale distribuée, et aussi le cholestérol est normale distribuée. On considère des données normale distribuées, quand le coefficient d' asymétrie et le coefficient d' aplatissement se trouvent entre [-1, 1] – voir le cours avec l' évaluation de la normalité des données. Donc Pearson est notre choix ici => Spearman est un réponse incorrecte
- c) Le coefficient de détermination est le carré du coefficient de corrélation Pearson = $-0,3 * -0,3 = 0,09$
- d) Oui, parce que le coefficient d' asymétrie est <0 (-0,3), qui indique une queue a la gauche
- e) L' interprétation classique du coefficient de détermination, ici égale a 9% (0,09 – voir en haut)

58

Exemples des questions pour l'examen

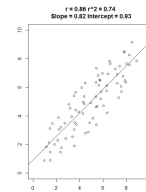
6) Lesquelles des réponses suivantes sont correctes en ce qui concerne la relation entre deux variables quantitatives montrée dans le graphique à cote:

- a) la relation semble être linéaire
- b) on peut calculer le coefficient de corrélation Pearson
- c) entre les variables il y a une relation proportionnelle
- d) entre les variables il y a une relation inversement proportionnelle
- e) le pente de la droite est négative

Réponse: a, b, c

Explications:

- a) On observe dans le graphique que le nuage des points semble suggérer qu'il y a une tendance que les points sont plutôt situés autour d'une droite imaginaire
- b) Dans l'énoncé on observe que les variables sont quantitatives. Si la forme du nuage des points semble être elliptique, ça indique que les deux variables sont normalement distribuées. En plus on a observé que la relation est plutôt linéaire. On n'a pas dans le scénario d'indications que les observations sont dépendantes, donc on considère que les paires des observations sont indépendantes. Donc on a toutes les conditions d'application du coefficient de corrélation Pearson
- c) On a une sensation visuelle que les points ont une tendance croissante (plus les valeurs de x augmentent, le y augmente aussi) => tendance croissante/ pente ascendante/ pente positive/ lien (direct) proportionnel
- d) Faux, voir c)
- e) Faux, voir c)



59

Exemples des questions pour l'examen

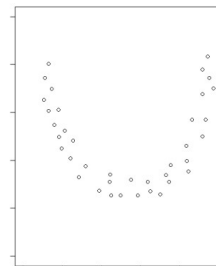
7) Lesquelles des réponses suivantes sont correctes en ce qui concerne la relation entre deux variables montrée dans le graphique à cote:

- a) la relation semble être non linéaire
- b) on ne peut pas calculer le coefficient de corrélation Pearson
- c) entre les variables il y a une relation proportionnelle
- d) entre les variables il y a une relation inversement proportionnelle
- e) le pente de la droite est négative

Réponse: a, b

Explications:

- a) On observe dans le graphique que le nuage des points NE semble pas suggérer qu'il y a une tendance que les points sont plutôt situés autour d'une droite imaginaire. Ici on observe une tendance courbe, plutôt indicative d'une relation quadratique
- b) Dans l'énoncé on observe que les variables sont quantitatives. Mais, on a observé que la relation n'est pas linéaire. Donc d'une des conditions d'application du coefficient de corrélation Pearson, n'est pas la => on ne peut pas l'utiliser
- c) On n'a pas une sensation visuelle que les points ont une tendance croissante (plus les valeurs de x augmentent, le y augmente aussi) Ici la relation n'est pas linéaire. Initialement il est une relation inversement proportionnelle, mais puis elle est proportionnelle. Donc la relation n'est pas purement proportionnelle
- d) Faux, voir c) Ici la relation n'est pas purement inversement proportionnelle
- e) Faux, voir d)



60

Exemples des questions pour l'examen

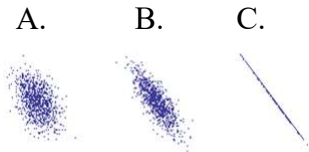
8) Lesquelles des réponses suivantes sont correctes en ce qui concerne la relation entre deux variables quantitatives montrée dans le graphique a cote:

- a) la relation entre les deux variables est plus forte pour le graphique B que pour le graphique A
- b) le coefficient de corrélation Pearson pour le graphique C est -1
- c) entre les variables il y a une relation proportionnelle
- d) entre les variables il y a une relation inversement proportionnelle
- e) le coefficient de corrélation Pearson pour le graphique B est plus grand en valeur absolue que pour le graphique A

Réponse: a, b, d, e

Explications:

- a) On observe que les points sont plus proche a une droite de régression au (« milieux ») des points dans le graphique B, que dans le graphique A. La pente des deux droites de régression, pour les deux images, semble être très similaire. Le degré de rapprochement plus grand, indique que l'intensité de la corrélation linéaire est plus grande.
- b) Touts les points sont sur une droit – ca indique une coefficient de corrélation de 1. Mais la pente de la droite est descendante, qui indique que le signe du coefficient de corrélation est négative => le coefficient de corrélation Pearson pour le graphique C est -1
- c) On observe que la tendance des toutes nouages des points est descendante, qui indique des relations inversement proportionnelles
- d) Voir c)
- e) Voir a) Ca indique que le coefficient de corrélation Pearson est plus distant de 0, dans le cas du B, que le cas du A. Mais le signe du coefficient est négatif. C'est pour cela on a la précision: en valeur absolue.



61

FLN...

L'hiver arrive ... proche de Cluj – 3 heures - se baigner dans des eaux thermales près de la neige – « **Baile Felix** » a trouver sur Google



Objectifs qui méritent une visite en Roumanie

62