GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=14271

Cours 10a

- Processus de recherche, études (*vidéo seulement*)
- Données bivariées
- Modélisation des données bivariées
- Régression linéaire simple
- Distribution de Y conditionnelle sur X
- Distribution d'échantillonnage des paramètres
- Régression multiple

Processus de recherche

- *Question* d'intérêt scientifique
- Décider : *quelles données* à recueillir (et comment)
- Collecte et *analyse* des données
- Conclusions, généralisations : *inférence* sur la population
- *Communication* et diffusion des résultats

Question Générique : Est-ce qu'un 'traitement' produit-il un 'effet' ?

Exemples :

- Fumer provoque-t-il le cancer, les maladies cardiaques, *etc* ?
- Est-ce que la consommation d'avoine diminue le taux de cholestérol ?
- L'échinacée prévient-elle le maladies ?
- Est-ce que l'exercice ralentit le processus de vieillissement ?

Genres d'études

- Une méthode simple pour résoudre ce type de question consiste à *comparer deux groupes* de sujets de l'étude :
  - *Groupe contrôle* : fournit une base de comparaison
  - *Groupe traitement* : groupe recevant le 'traitement'
- *Étude expérimentale* : sujets affectés aux groupes (traitement, contrôle) par l'investigateur
  - *randomisation* : protège contre les biais dans l'attribution aux groupes
  - *'aveugle', 'double-aveugle'* : protège contre les biais dans l'évaluation des résultats
  - *placebo* : traitement artificiel
- *Étude d'observation* : sujets 'attribuent' eux-mêmes aux groupes
  - *facteur de confusion* : un facteur qui présente une association avec le facteur de risque examiné *et* avec le résultat

Commentaires

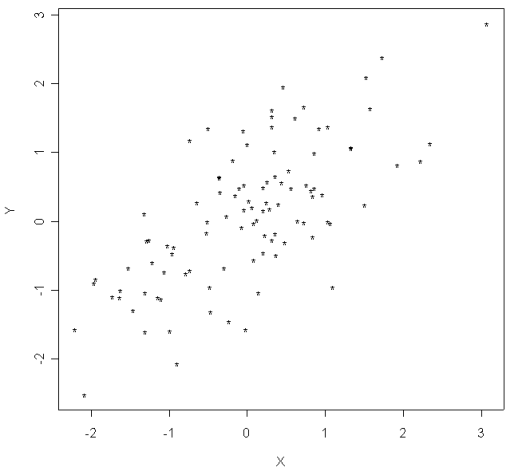
- Avec une expérience contrôlée bien planifiée et exécutée, il est possible de déduire *la causalité*
- Ceci *n'est pas possible* avec les études d'observation en raison de la présence de facteurs de confusion
- En présence de facteurs de confusion, il n'est pas possible de dire si la différence observée entre les groupes est due au *traitement* ou au *facteur de confusion*
- Pas toujours possible de mener une étude expérimentale, pour des raisons *pratiques* et *éthiques*

Données bivariées

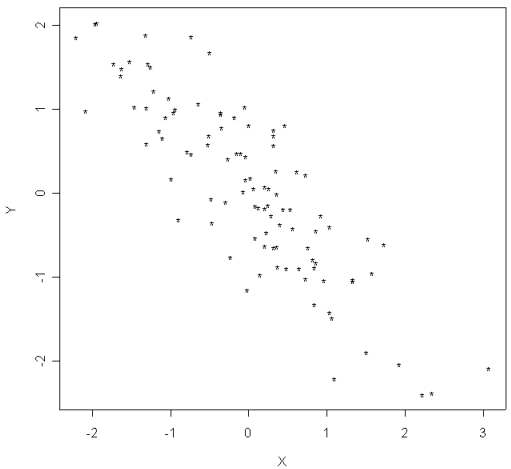
- Mesures de *deux* variables ; p. ex. X et Y
- On considère le cas de deux variables *continues*
- On veut découvrir la *relation* entre les deux variables
  - longueur de l'avant-bras et taille
  - taille et poids
  - expressions de gène A et gène B
- Considérons les ensembles de données qui sont (au moins approximativement) *normales bivariées* $\Leftrightarrow$ forme ovale
- $(X, Y) \sim BVN((\mu_x, \mu_y), (\sigma_x^2, \sigma_y^2), \rho)$

Analyse exploratoire : Diagramme de dispersion

- **Résumé graphique** d'un jeu de données bivariées à l'aide d'un *diagramme* (ou *nuage*) *de dispersion*
- Les valeurs d'une variable sur l'axe horizontal et les valeurs de l'autre sur l'axe vertical
- Peut être utilisé pour voir comment les valeurs de 2 variables tendent à évoluer les unes avec les autres (c'est-à-dire comment les variables sont **associées**)

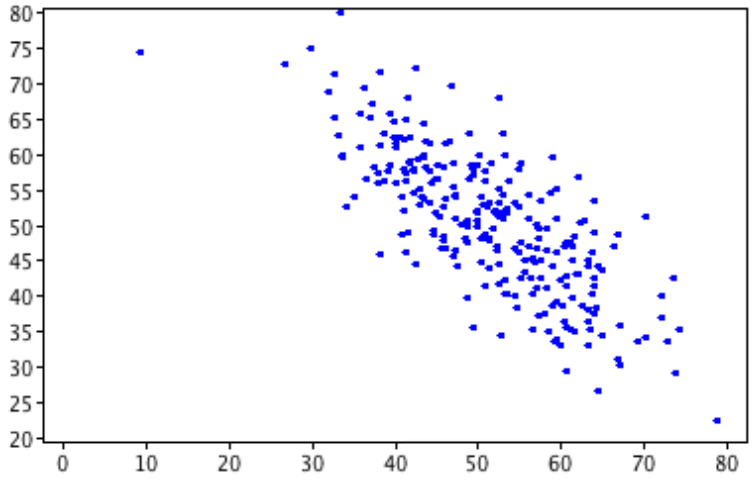


(a) association positive

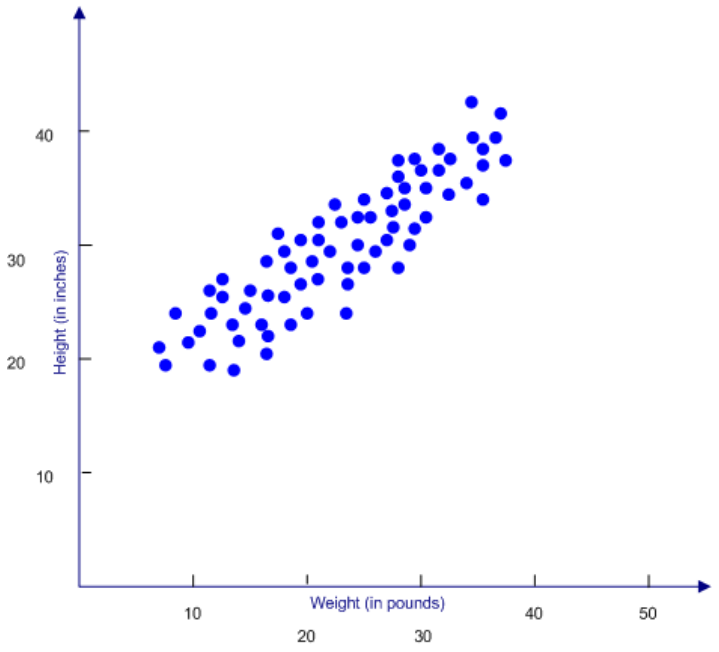


(b) association négative

Diagramme (nuage) de dispersion



(a)



(b)

QCM : *Quelle est l'association entre X et Y ? ?*

(a) nulle **(b)** positive **(c)** négative **(d)** impossible à déterminer

Figure (a) : _____ Figure (b) : _____

Résumés numeriques

- Typiquement, les données bivariées sont résumées (numériquement) avec **5 statistiques**
- Celles-ci fournissent un bon résumé pour les nuages de points avec la même forme générale que nous venons de voir (ovale)
- On peut résumer chaque variable *séparément* : $\overline{X}, s_X; \overline{Y}, x_Y$
- Mais ces valeurs ne disent pas comment les valeurs de X et Y *varient ensemble*

Corrélation

- Soient X et Y VAs, et $Var(X) > 0, Var(Y) > 0$
- $Cov(X,Y) = E[(X - EX) \times (Y - EY)]$ $(= E(XY) - EX \times EY)$
- La **corrélation** $\rho(X, Y)$ est définie ainsi :

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

- ρ est une *quantité sans unités*, $-1 \leq \rho \leq 1$
- La corrélation ρ, comme la covariance, est **une mesure d'association linéaire** (le degré de linéarité) des VAs X et Y
- Les valeurs ρ proches de 1 ou -1 indiquent une linéarité quasiment rigoueuse entre X et Y, tandis que des valeurs proches de 0 indique une absence de toute relation **linéaire**
- Le signe de ρ indique la direction de l'association (positive ou négative, correspondant à la pente de la droite)
- Lorsque $\rho(X, Y) = 0$, X et Y sont **non-corrélées**

Coefficient de corrélation de l'échantillon

- Le **coefficient de corrélation de l'échantillon** r (ou $\hat{\rho}$) est défini comme la valeur moyenne du produit (normalisé) XY :

$$r = E[(X \text{ centrée-réduite}) * (Y \text{ centrée-réduite})]$$

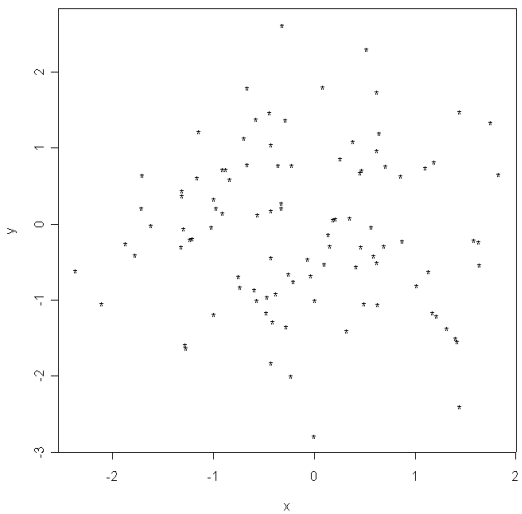
- centrée-réduite $=$ standardisée (normalisée)
  $= (X - \text{ moyenne}(X))/\text{écart-type}(X)$
- r est une quantité *sans unités*
- $-1 \leq r \leq 1$
- r est une mesure d'**ASSOCIATION** **LINÉAIRE**

Corrélation ≠ Causalité

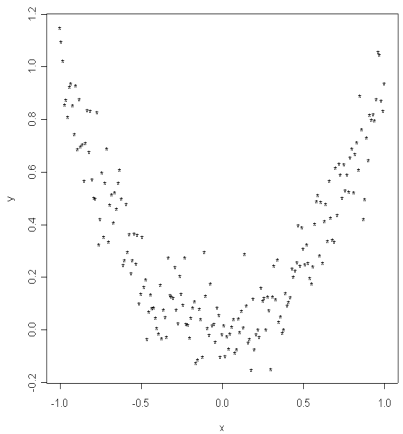
- On *ne peut pas en déduire* que, puisque X et Y sont fortement corrélées (r proche de -1 ou 1) que X est *à l'origine* (ou la *cause*) d'un changement dans Y
- Y pourrait être la cause de X
- X et Y les deux pourraient varier avec un tiers, un facteur peut-être inconnu (soit de causalité ou pas, souvent le temps)
  - polio et boissons non alcoolisées
  - nombre de pompiers envoyés à un incendie et montant des dégâts
  - Les enfants qui reçoivent un soutien scolaire obtiennent de moins bonnes notes que ceux qui ne le reçoivent pas
- Si $r \approx 0$, il n'y a pas d'**ASSOCIATION** **LINÉAIRE**

  – ceci n'est **PAS** à dire qu'il n'y a *AUCUNE ASSOCIATION*
- On ne peut pas en déduire la forme du diagramme de dispersion seulement à partir de la valeur de r

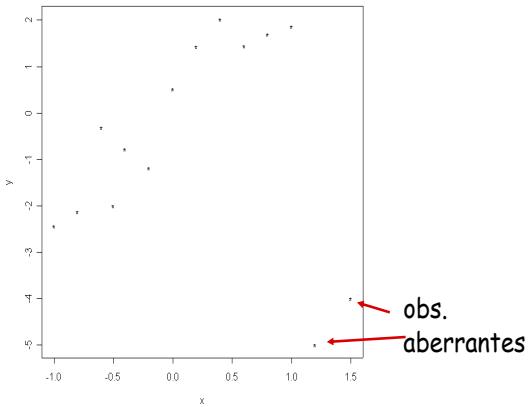
$r \approx 0$



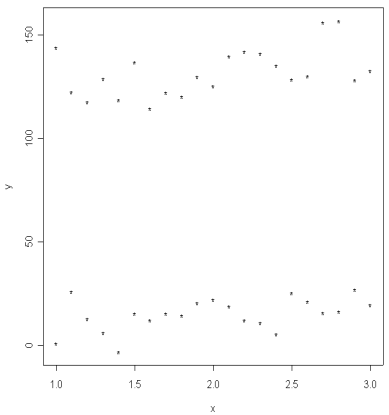
(a) dispersion au hasard



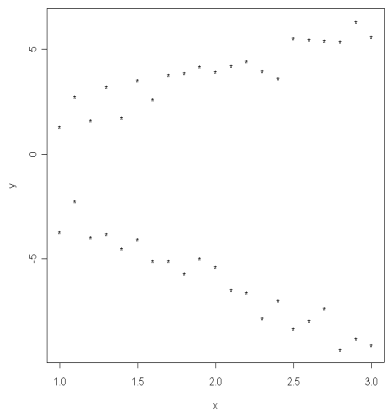
(b) courbe



(c) observations aberrantes



(d) parallélisme

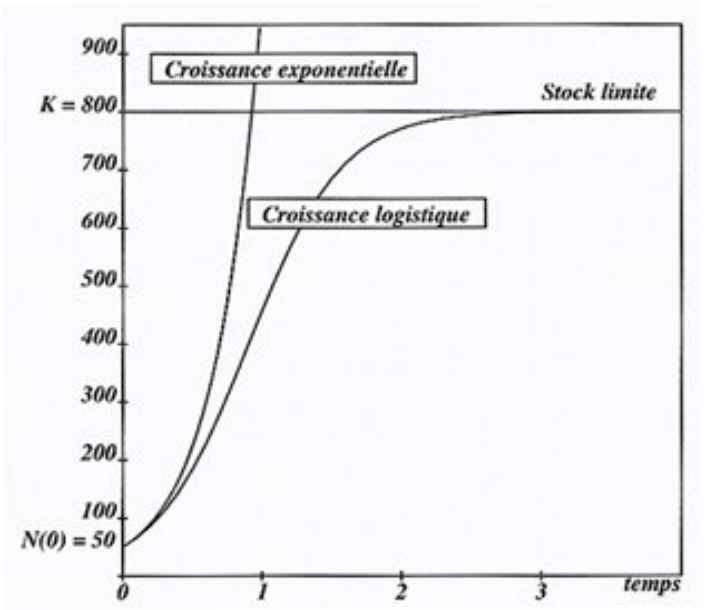


(e) deux droites différentes

Modélisation d'un nuage de forme ovale

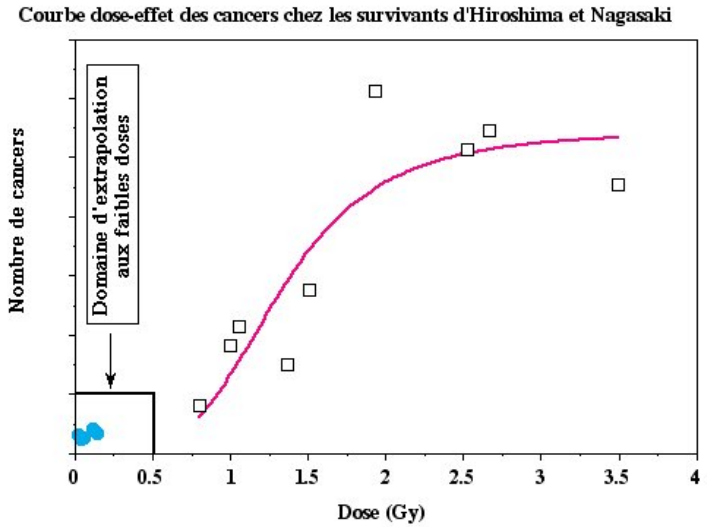
- Variable à expliquer / variable réponse : Y
- Variable explicatrice / prédictrice : X
  - La valeur de X est supposée connue *sans erreur*
  - On suppose que les variations de Y sont *influencées* par X
  - Le modèle permet d'exprimer sous la forme d'une *relation mathématique* la liaison supposée
- La connaissance de ces variables permettent à l'aide du modèle de *prédire* Y
  - Estimater les valeurs de Y :
    – *ponctuellement*
    – par *intervalle*
- Le modèle permet de mesurer *l'impact* (ou *l'effet)* d'une variable explicative sur Y

Relation déterministe ou statistique



(a) déterministe

- *Une seule valeur* de Y pour une valeur de X



(b) statistique

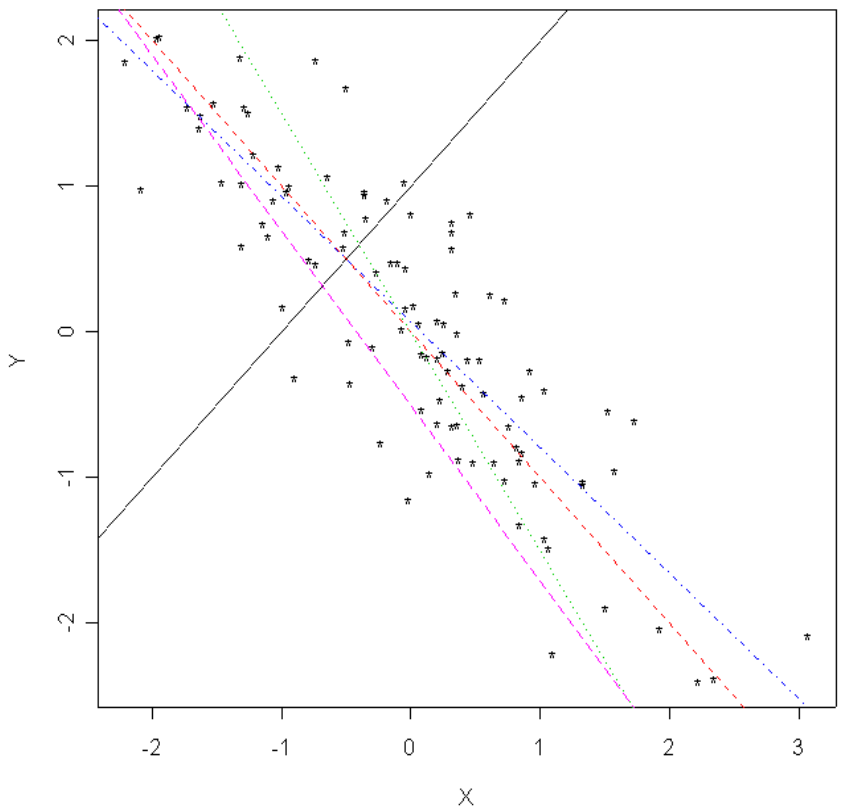
- *Plusieurs valeurs* de Y pour une valeur de X
- *'Probabiliser' Y* pour une valeur fixe de X

Régression linéaire simple

- Se réfère à tracer une droite (particulière) à travers un nuage de points
- Utilisé pour les 2 objectifs :
  - Explication
  - Prévision
- Modèle linéaire statistique :
  - $Y = \beta_0 + \beta_1 X + \epsilon \quad \Rightarrow \quad E[Y \mid X] = \beta_0 + \beta_1 X$
  - $E(\epsilon) = 0; \;\; Var(\epsilon) = \sigma^2$
- L'équation d'une droite de prédire Y quand on connaît la valeur spécifique x : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\beta_0 =$ l'*ordonnée à l'origine*; $\beta_1 =$ la *pente* (dans la population)

Quelle droite ?

- Il y a beaucoup de droites qui pourraient être faites à travers le nuage de points
- Comment choisir ?



Prévision par régression

- On peut faire une prévision en utilisant *la droite de régression* :

  lorsque X augmente de 1 (écart-type), la valeur prédite Y augmente ** PAS de 1 (écart-type) **,
  mais seulement de r (écart-type) (vers le bas si r est négatif) :

  - $$\frac{\hat{Y} - \overline{Y}}{s_Y} = \boldsymbol{r} \frac{X - \overline{X}}{s_X}$$

- Cette prévision pourrait s'exprime également dans la forme :

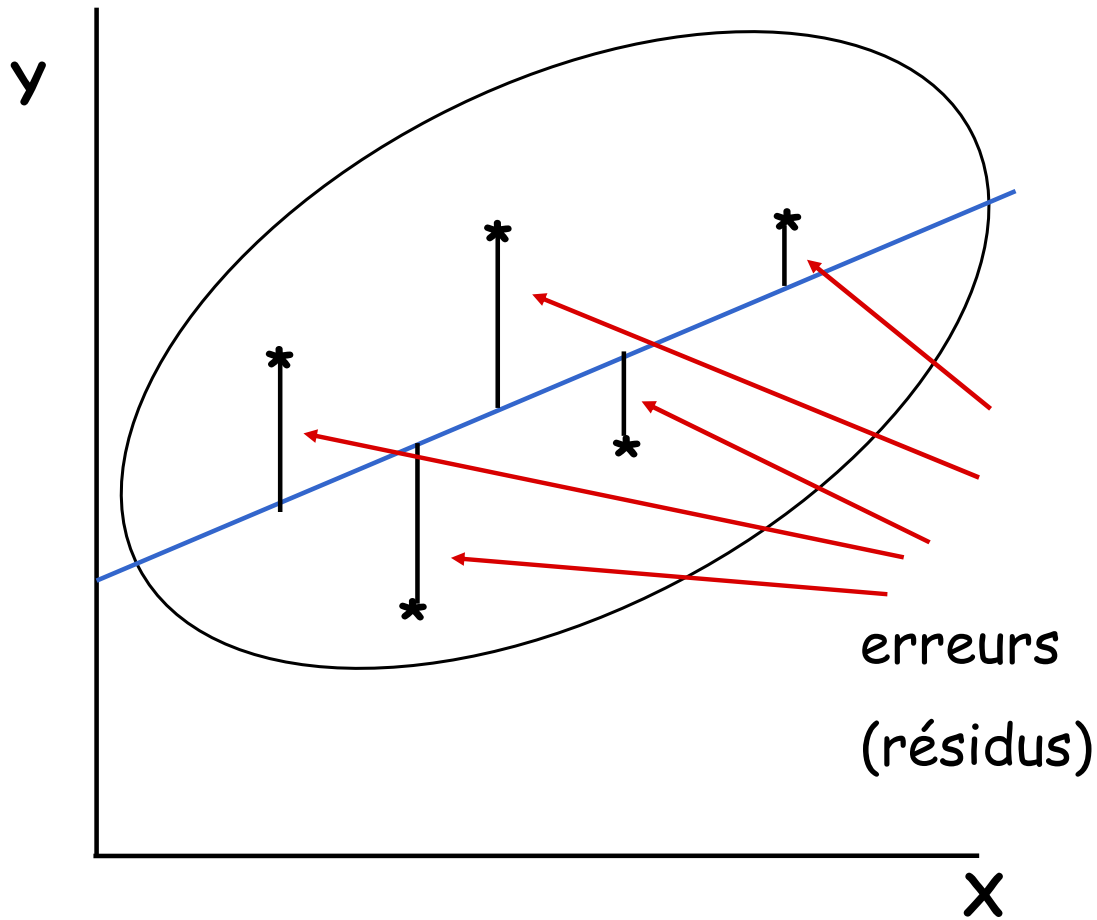
  $$\text{préd. } y = \text{ord.} + \text{pente} \times x, \text{ avec}$$

  - pente $= \hat{\beta}_1 = r\, s_Y / s_X$
  - ord. $= \hat{\beta}_0 = \overline{y} -$ pente $\times\, \overline{x}$

Moindres carrés

Q : D'où vient cette équation ?

R : C'est la droite qui est 'meilleure' dans le sens que la somme des *carrés des erreurs* dans le plan vertical (Y) est au *minimum*



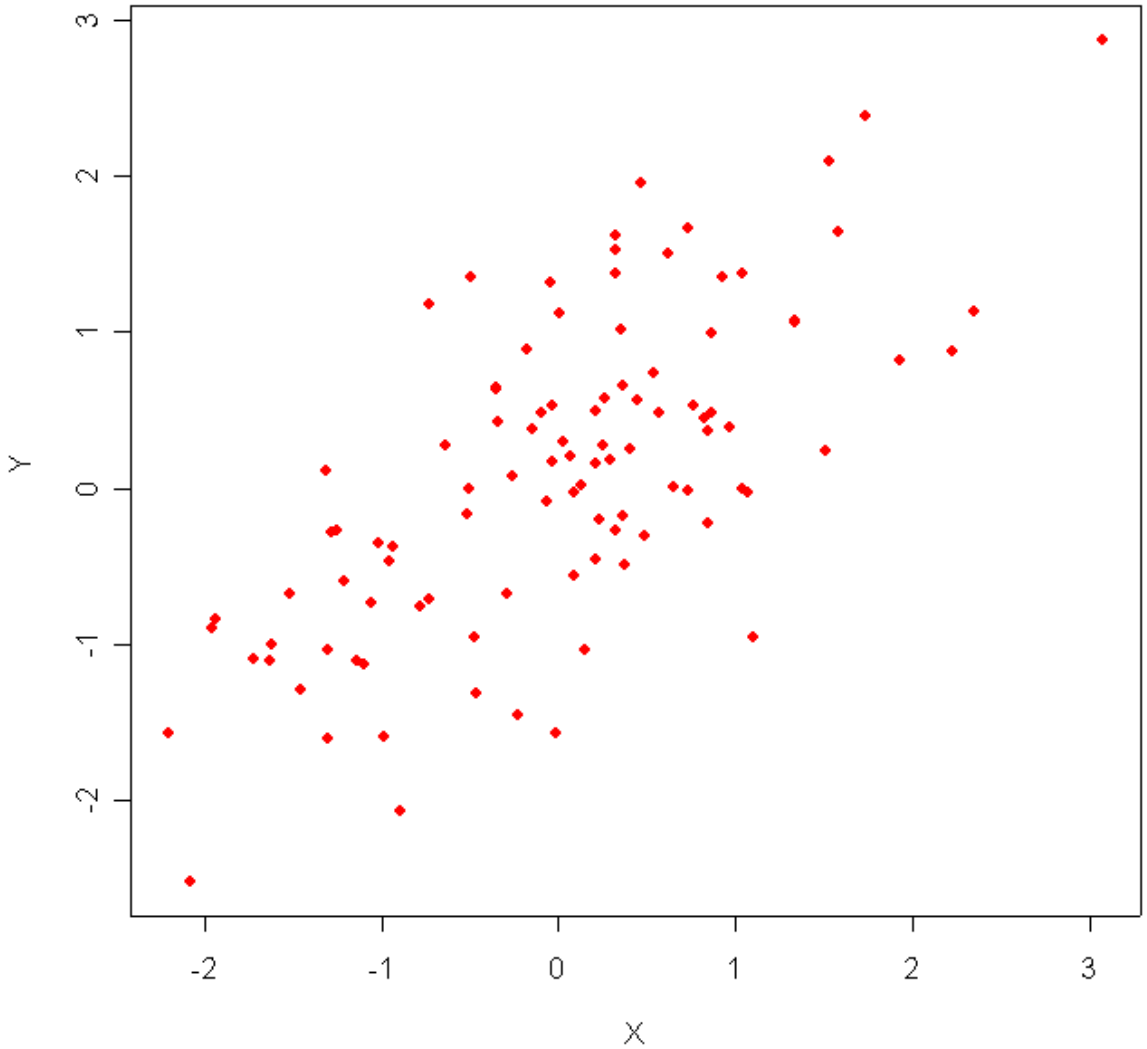
Interprétation des paramètres

- L'équation de droite de régression comprend 2 paramètres : la *pente* et l'*ordonnée à l'origine*
- La *pente* est le changement moyen de Y pour un changement de X de 1 unité
- L'*ordonnée à l'origine* est la valeur de Y estimée lorsque $X = 0$
- Si la pente $= 0$, alors X n'aide pas à prédire Y (prédiction linéaire)

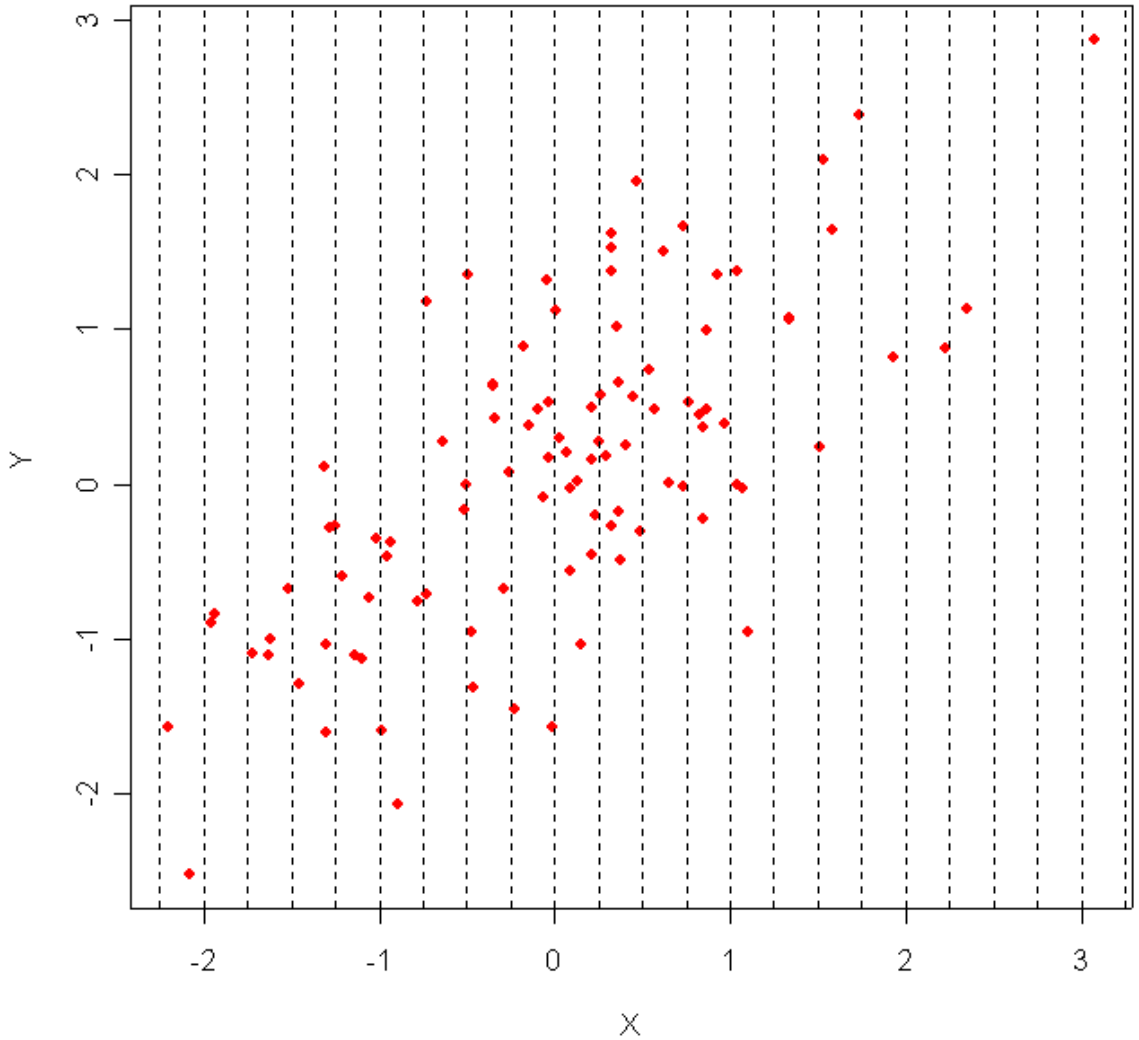
Une autre vue de la droite de régression

- On peut diviser le nuage de points dans les régions (*X-bandes*) fondées sur des valeurs de X
- Au sein de chaque X-bande, mettez la valeur moyenne de Y (en utilisant uniquement les valeurs de Y possèdant des valeurs X dans le X-bande)
- Il s'agit de la *courbe des moyennes*
- La droite de régression pourrait être considérée comme une *version lissée* de la courbe des moyennes

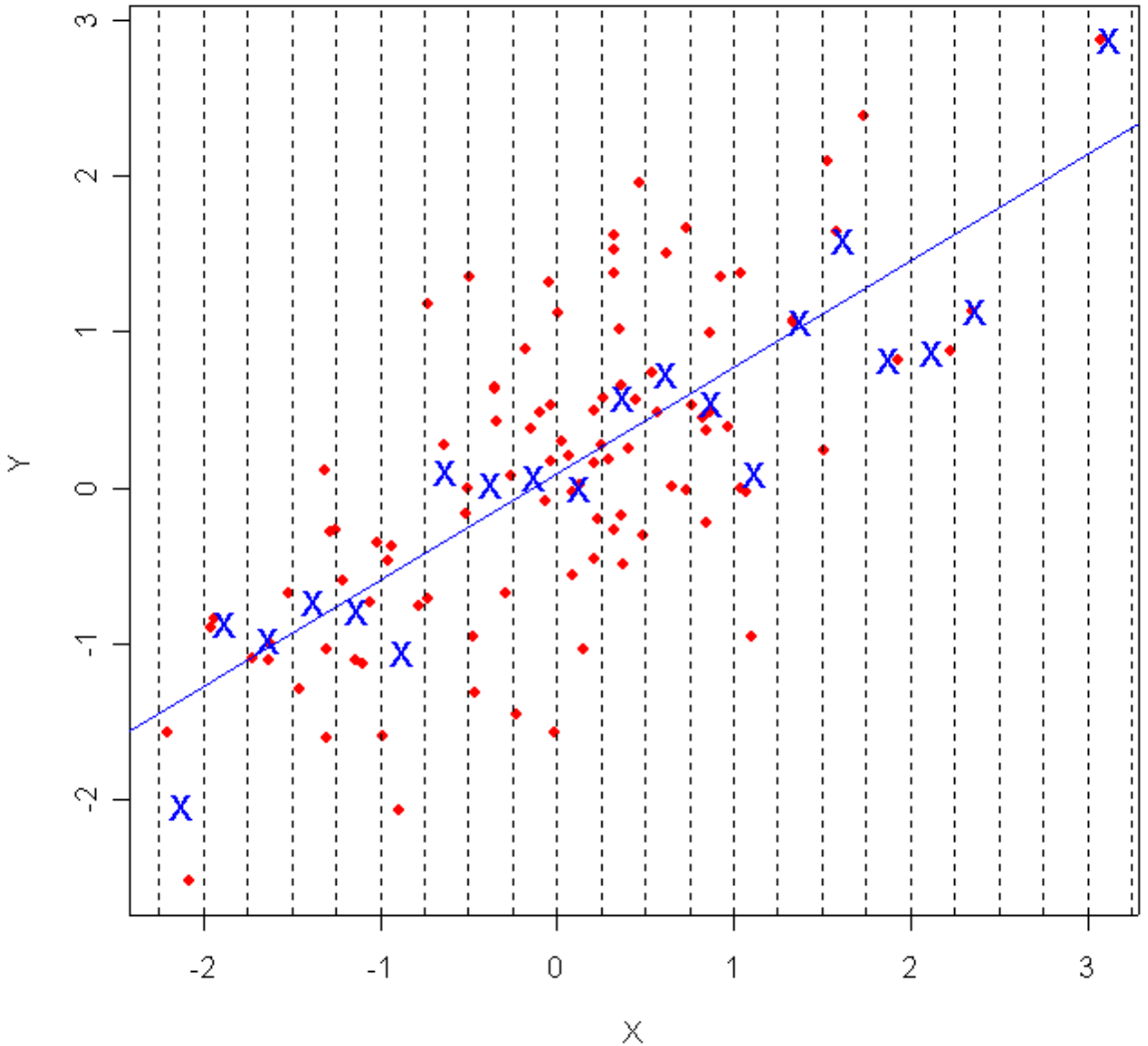
Diagramme de dispersion (encore une fois)



Création des X-bandes



Graphique des moyennes



Erreur quadratique moyenne (EQM) et l'erreur-type

- L'erreur-type : Il s'agit d'une nouvelle variabilité : la variabilité de l'espérance conditionnelle de Y sachant $X =$ l'*erreur-type*
- C'est la racine carrée de l'erreur quadratique moyenne : $EQM =$ la moyenne arithmétique* des carrés des écarts entre les prévisions et les observations
- *(au lieu de division par n, diviser par le nombre de degrés de liberté (df))
- $REQM(Y) = s_Y\sqrt{(1-r^2)}$

PAUSE

Démarche de la régression

A partir d'un échantillon de valeurs pour la variable réponse Y et la (ou les) variables prédictrices X :

- Vérifier la possibilité d'une liaison linéaire entre Y et X
  - représentation graphique
  - coefficient de correlation
- Estimation des paramètres
  - coefficients $\beta_i \Rightarrow \hat{\beta}_i$
  - écart-type pour les erreurs $\sigma \Rightarrow \hat{\sigma}$
- Evaluation du modèle (la semaine prochaine)
  - indices de qualité R^2, R^2_{aj}
  - évaluation globale de l'ajustement (F de Fisher)
  - test(s) de coefficients individuellement
  - étude des résidus, détection des points abérrants, influentiels

Résumé : Régression linéaire simple (conceptuelle)

- Pour un diagramme de dispersion qui est *de forme ovale*, nous pouvons trouver une droite qui sert à résumer les points
- Un principe souvent utilisé pour l'ajustement de cette droite est *moindres carrés* : le total des carrés des erreurs (verticales) est réduit au minimum
- Selon ce principe, la prédiction de régression pour Y sachant X nous dit que :
  lorsque X augmente de 1 fois l'écart-type, Y (en espérance) augment de r fois l'écart-type
- On peut trouver l'équation de la droite des moindres carrés en utilisant les 5 statistiques :

$$\overline{X}, SD(X), \overline{Y}, SD(Y), r$$

- La *pente* (estimée) égale à $\hat{\beta}_1 = r\dfrac{s_Y}{s_X}$,
  l'*ordonnée à l'origine* (estimée) est $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

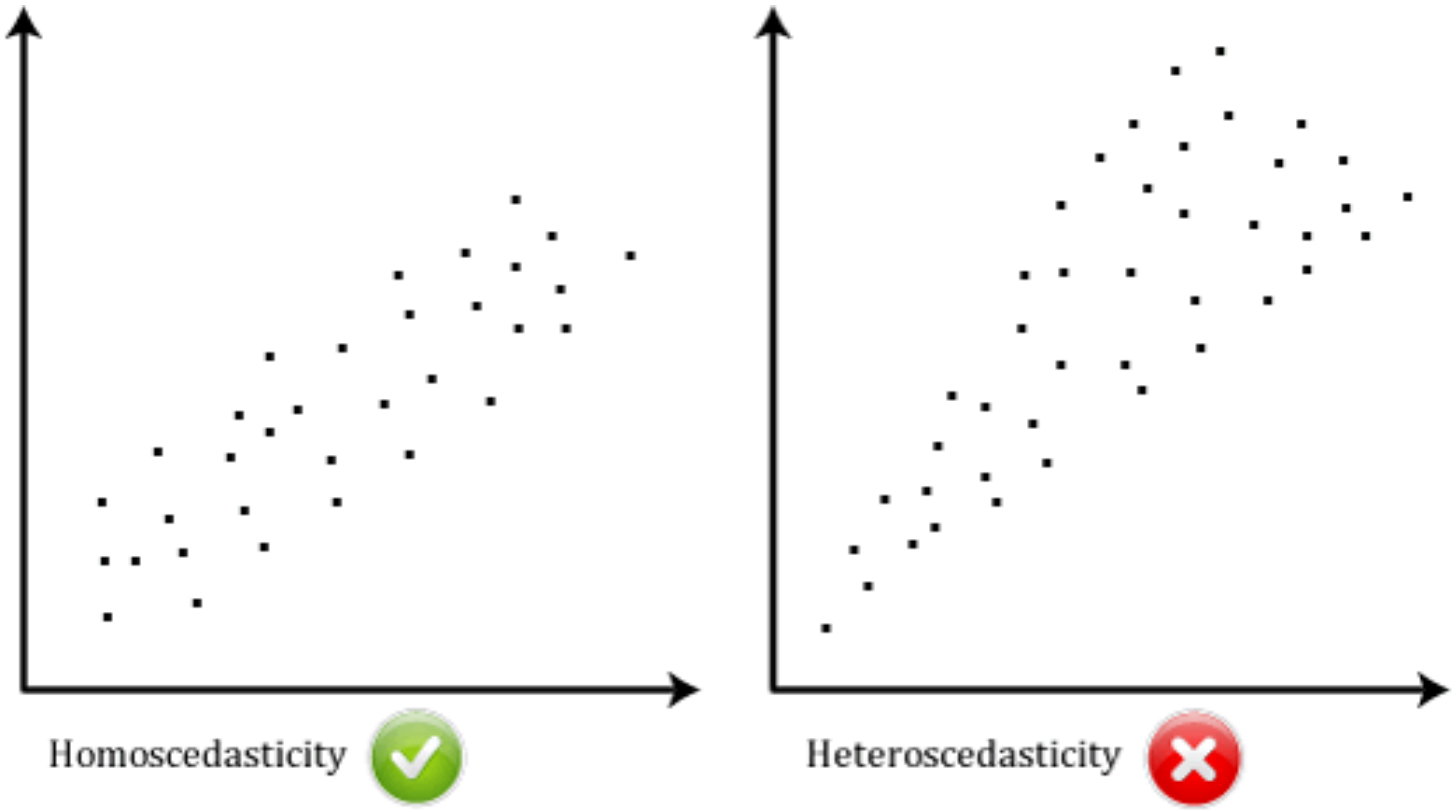
Régression linéaire simple – cadre mathématique

- Ici, on considère un modèle où la *variable expliquée* (ou *réponse*) y_i a une association linéaire à une *variable explicative* (ou *régresseur* ou *prédictrice*) x_i :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

- $\epsilon_1, \ldots, \epsilon_n$ sont supposés variables aléatoires
  - non corrélées
  - espérance $= 0$
  - variance $= \sigma^2$ pour tout $i = 1, \ldots, n$ (*homoscédastique*)
- x_i sont supposés être des constantes (mesurés sans erreur)
- $\Longrightarrow$ Si les erreurs sont aussi supposées *normalement distribuées*, on peut faire les *tests* et les *intervalles de confiance (IC)*

Erreurs homoscédastiques, heteroscédastiques



Méthode des moindres carrés

- **Ces détails ne seront pas examinés ! !**
- Les données ne sont qu'un *échantillon* (et ne sont pas l'ensemble de la population)
- Donc il faut *estimer* les valeurs des paramètres β_0 (ordonnée à l'origine) et β_1 (pente) (également la variance des erreurs σ^2) :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$$

- Selon le *principe des moindres carrés*, on cherche les estimateurs qui réduisent au minimum :

$$SC(\hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- ('SC' = 'somme des carrés' = 'sum of squares' en anglais)

Méthode de moindres carrés, cont.

C'est maintenant *un problème d'optimisation*, de trouver des valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$ qui réduisent au minimum

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

Pour résoudre ceci, dériver par rapport à β_0, β_1 ; trouver les zéros :

$$\begin{aligned}
\frac{d}{d\beta_0} &= \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \qquad (*)
\end{aligned}$$

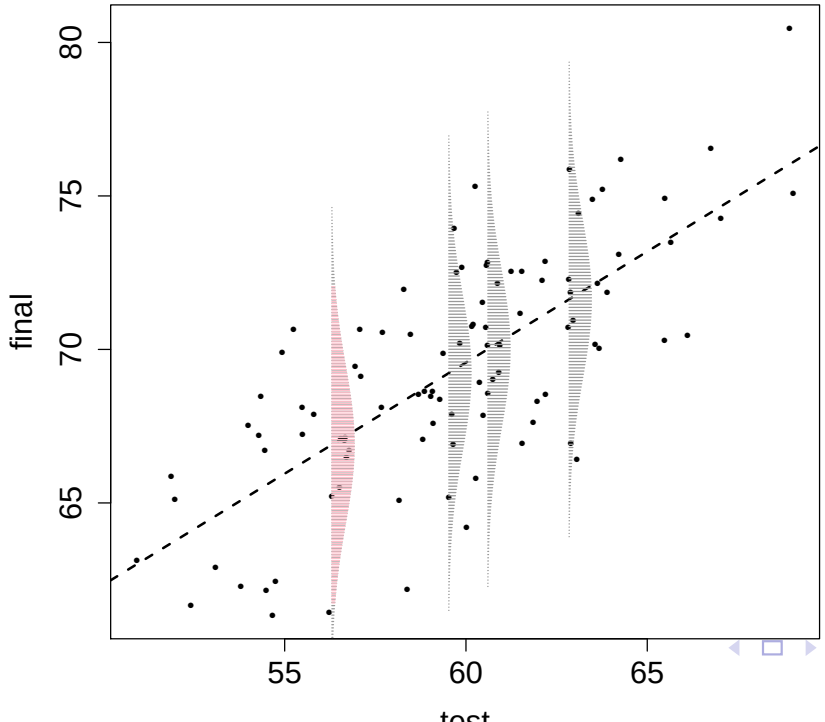
Moindres carrés, cont.

$$
\begin{aligned}
\frac{d}{d\beta_1} &= \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) = 0 \\
&\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
&\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \qquad (**)
\end{aligned}
$$

Solution simultanée de (*) et de (**) pour les paramètres β_0 et β_1 nous donne **l'estimation de régression**.

Distribution normale conditionelle

- L'*espérance* est la prévision $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- L'*erreur-type* ($REQM$) est la racine carrée de l'erreur quadratique moyenne : EQM = la moyenne arithmétique* des carrés des écarts entre les prévisions et les observations
- *(au lieu de division par n, diviser par le nombre de df)
- $REQM(Y) = s_Y \sqrt{(1 - r^2)}$



Exemple

Exemple 10a.1 Pour un échantillon aléatoire d'hommes d'age 25–34 ans, la relation entre la taille et les revenus pourrait être résumée par les cinq statistiques (vous pouvez supposer que le nuage de points est de *forme ovale*) :

taille moyenne $= 70$ pouces ; $s_T = 3$ pouces
revenu moyen $= \$29{,}800$; $s_R = \$14{,}400$; $r = 0.2$.

(a) Environ quel pourcentage des hommes ont un revenu supérieur à \$ 37,000 ?

(b) Parmi les hommes qui sont de 73 pouces de taille, environ quel pourcentage des revenus sont supérieurs à \$37,000 ?

Propriétés de l'estimateur de la pente

- L'estimation de la droite de régression : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- L'estimateur des moindres carrés pour la pente β_1 pourrait être écrit comme :

$$\hat{\beta}_1 = \frac{y_1 (x_1 - \overline{x}) + \cdots + y_n (x_n - \overline{x})}{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}$$

- L'espérance de l'estimateur : $E[\hat{\beta}_1] = \beta_1$
- La variance de l'estimateur :

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}$$

- Il nous faut un estimateur de σ^2 $(e_i = y_i - \hat{y}_i)$:

$$\hat{\sigma}^2 = \frac{e_1^2 + \cdots + e_n^2}{n - 2}$$

Test/Intervalle de confiance pour la pente

- Pour tester $H : \beta_1 = \beta_1^H$ contre $A : \beta_1 \neq \beta_1^H$:

$$t\text{-pente}_{obs} = \frac{\hat{\beta}_1 - \beta_1^H}{\hat{\sigma}/\sqrt{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}}$$

- On REJETTE H si : $|t\text{-pente}_{obs}| > t_{\underline{n-2},1-\alpha/2}$
- Le IC de niveau $1-\alpha$ pour la pente β_1 est :

$$\hat{\beta}_1 \pm \frac{\hat{\sigma}}{\sqrt{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}}\, t_{\underline{n-2},1-\alpha/2}$$

Données multivariées

Individus	X_1	X_2	$\ldots$	X_j	$\ldots$	X_p
i_1	x_{11}	x_{12}	$\ldots$	x_{1j}	$\ldots$	x_{1p}
i_2	x_{21}	x_{22}	$\ldots$	x_{2j}	$\ldots$	x_{2p}
$\ldots$						
i_i	x_{i1}	x_{i2}	$\ldots$	x_{ij}	$\ldots$	x_{ip}
$\ldots$						
i_n	x_{n1}	x_{n2}	$\ldots$	x_{nj}	$\ldots$	x_{np}

vecteur des moyennes : $(\overline{x}_1, \ldots, \overline{x}_p)$
matrice des variances-covariances (ou *matrice de dispersion*) :

$$\begin{pmatrix} s_1^2 & s_{1,2} & \cdots & s_{1,p} \\ s_{2,1} & s_2^2 & \cdots & s_{2,p} \\ \cdots & s_i^2 & s_{i,j} & \cdots \\ s_{p,1} & s_{p,2} & \cdots & s_p^2 \end{pmatrix}$$

Logiciel : *R*

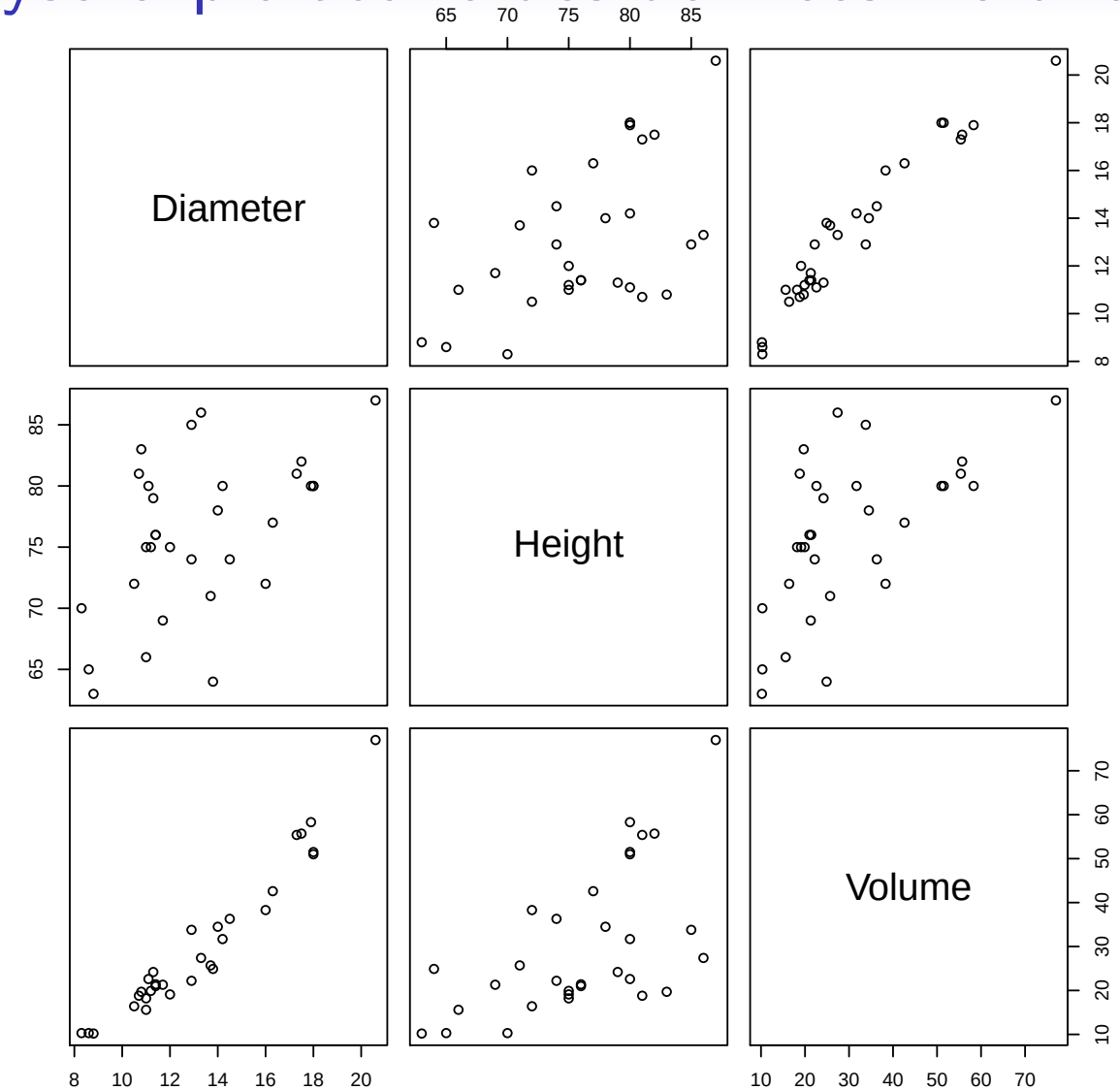
Pourquoi *R* ?

- Puissant, flexible, extensible langue et environnement pour le calcul statistique
- Large gamme de fonctions statistiques intégrées et 'packages' disponibles
- De haute qualité, des capacités graphiques excellentes
- Disponible pour les systèmes Unix / Linux, Windows, Mac
- Tout cela et ... *R* est gratuit !
- `http://cran.r-project.org/`

Exemple

- Un échantillon de cerisiers a été coupé et les mesures prises pour
  - `Diameter` (inches)
  - `Height` (feet)
  - `Volume` (cubic feet)
- Le but de la collecte de ces données était de fournir un moyen de prédire le volume de bois dans les arbres, sachant la hauteur et le diamètre
- Utilise un modèle de régression

Analyse exploratoire des données multivariées



Algèbre matricielle pour la régression

- Le modèle :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Régression multiple

- On peut avoir *plusieurs* variables explicatives x :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- Même suppositions dans le cas régression simple : $\epsilon \sim$ iid $N(0, \sigma^2)$
- $E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$
- Solution moindres carrés :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

où $\boldsymbol{X}$ est la *matrice d'expérience* (*design matrix*)

Moindres carrés (ordinaires) pour la régression multiple

- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Trouver une solution $\hat{\boldsymbol{\beta}}$ qui minimise la somme des carrés des résidus (solution de *moindres carrés ordinaires (MCO)*) :

$$\min \sum_{i=1}^{n} e_i^2 \implies \frac{\partial \left(\sum_{i=1}^{n} e_i^2 \right)}{\partial \hat{\beta}_j} = 0, \quad j = 0, \ldots, p$$

$$\implies \sum_{i=1}^{n} x_{ij} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0, \quad j = 0, \ldots, p$$

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0 \implies \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

$$\implies \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

où $\boldsymbol{X}$ est la *matrice d'expérience* (*design matrix*) et $\boldsymbol{X}'$ est la transposée de $\boldsymbol{X}$

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

équation → y : x_1 : x_2

$\hat{\beta}_0$ → (Intercept) $\hat{\beta}_1$ → Diameter $\hat{\beta}_2$ → Height

Volume = -57.99 + 4.71 x Diameter + 0.34 x Height

*** Interprétation des coefficients ***

- Les coefficients de régression correspondent aux changements anticipés dans la réponse lorsqu'un changement d'une unité survient dans une variable explicative/prédictrice
- Pour la régression simple :
  - la pente est le changement espéré de la variable réponse si la variable explicative (x) est augmentée de 1 unité
  - l'ordonnée à l'origine est la valeur prédite de la réponse (y) lorsque $x = 0$
- Une distinction très importante – lorsque l'équation comporte *plusieurs* variables prédictrices :
  - chaque coefficient $\beta_1, \ldots, \beta_p$ correspond à la contribution d'une variable lorsque **toutes les autres variables présentes dans l'équation sont contrôlées ou tenues constantes**
  - le coefficient β_0 est la valeur prédite de la réponse (y) lorsque **toutes les variables** $x_1, \ldots, x_p = 0$

Propriétés de l'estimateur MCO

Dans le cas

1. $E(\epsilon_i) = 0,\ i = 1, \ldots, n$;
2. $Var(\epsilon_i) = \sigma^2$ (constante) ;
3. $Cov(\epsilon_i, \epsilon_j) = Cor(\epsilon_i, \epsilon_j) = 0,\ i \neq j$

on a :

- Espérance : $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- Variance : $Var(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$
  $((\boldsymbol{X}'\boldsymbol{X})$ symmétrique)
- 0ptimalité :
  - Le théorème **Gauss-Markov** nous dit que *parmi toute estimation linéaire non biaisée*, l'estimateur (MCO) possède la *variance minimale*
  - On peut le résumé en disant : **l'estimateur MCO est le « BLUE »** (Best Linear Unbiased Estimator)

Test/intervalle de confiance pour les coefficients

- En supposant en plus $\epsilon_1, \ldots, \epsilon_n \sim$ iid $N(0, \sigma^2)$, on a

$$\hat{\boldsymbol{\beta}} \sim MVN\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$$

- Donc, $Var(\hat{\beta}_i) = \sigma^2 \left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{i+1,\, i+1}$
- L'IC avec indice de confiance $1 - \alpha$ pour β_i prend la forme

$$\hat{\beta}_i \pm \hat{\sigma}\sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{i+1,\, i+1}}\; t_{n-p-1,1-\alpha/2}$$

- Pour tester $H : \beta_i = 0$ contre $A : \beta_i \neq 0$

$$t_{obs} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{i+1,\, i+1}}}$$

- On REJETTE H si : $|t_{obs}| > t_{n-p-1,1-\alpha/2}$
(également si l'IC ne contient pas la valeur 0)

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```