h e g

Haute école de gestion de Genève
Geneva School of Business Administration

Statistique inférentielle

– Statistiques III –

Dr Christophe Hebeisen
christophe.hebeisen@hesge.ch

Cours 12 : régression linéaire simple

HEG - Économie d'Entreprise

automne 2010

Objectif

Savoir valider et estimer la qualité d'une régression linéaire simple.

Rappels

Test d'ajustement du χ^2

Supposition nécessaire pour certain tests : la population suit une loi spécifique (p.ex. loi normale).

1. Acquisition d'un échantillon de taille suffisamment grande.
2. Classement des données en k différentes catégories.
3. Calcul des fréquences absolues observées o_i.
4. Comparaison des fréquences absolues théoriques e_i et des fréquences observées o_i.

Hypothèse nulle considérée : H_0 = distribution théorique.

Critère basé sur un test du χ^2 : comme d'habitude, H_0 sera rejetée si une trop grande différence existe. On procédera « comme pour un test à droite » ($\chi^2 > \chi^2_{\alpha;n}$).

Statistique de test

La statistique à utiliser suit une loi du χ^2 à $k-1$ degrés de liberté et est calculée ainsi :

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

où

o_i = fréquence observée pour la catégorie i
e_i = fréquence théorique pour la catégorie i
k = nombre de catégories

Supposition : LA TAILLE DE L'ÉCHANTILLON EST SUFFISAMMENT GRANDE (en pratique, $n \geq 30$, taille par cellule ≥ 5, sinon regrouper).

Test d'indépendance

But : Tester l'indépendance de deux variables catégorielles.

Moyen : Table de contingence.

Principe : comparer les fréquences observées o_{ij} avec les fréquences théoriques e_{ij} en cas d'indépendance :

$$e_{ij} = \frac{(\text{Total ligne } i) \cdot (\text{Total colonne } j)}{\text{Taille de l'échantillon}} = \frac{n_{i.} \cdot n_{.j}}{n}$$

Hypothèse nulle considérée : $H_0 =$ les deux variables sont indépendantes.

Critère : l'hypothèse d'indépendance des variables sera rejetée « comme dans le cas d'un test à droite ».

Statistique du test

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

où

o_i = fréquence observée de la cellule (i, j)
e_i = fréquence théorique de la cellule (i, j)
r = nombre de lignes
c = nombre de colonnes

Supposition : LA TAILLE DE L'ÉCHANTILLON EST SUFFISAMMENT GRANDE.

En pratique, $n \geq 30$ et effectif observé par cellule ≥ 5.

Corrélation et régression linéaire

Coefficient de corrélation

Pour des données bivariées, le coefficient de corrélation linéaire de Pearson est défini par la covariance standardisée des deux variables, et peut se calculer ainsi :

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right) \cdot \left(\sum y_i^2 - n\bar{y}^2\right)}}$$

Une corrélation linéaire étant le plus souvent calculée à partir d'échantillons, sa valeur est sujette à des erreur d'échantillonnage. Ainsi, r_{xy} n'est qu'une estimation de la véritable valeur du coefficient de corrélation linéaire ρ.

Propriétés et interprétation

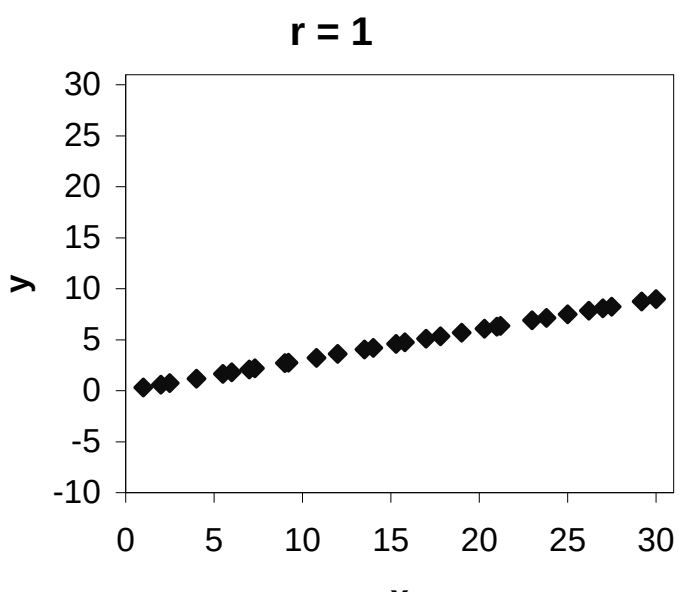
Le coefficient de corrélation linéaire est une grandeur comprise entre -1 et $+1$:

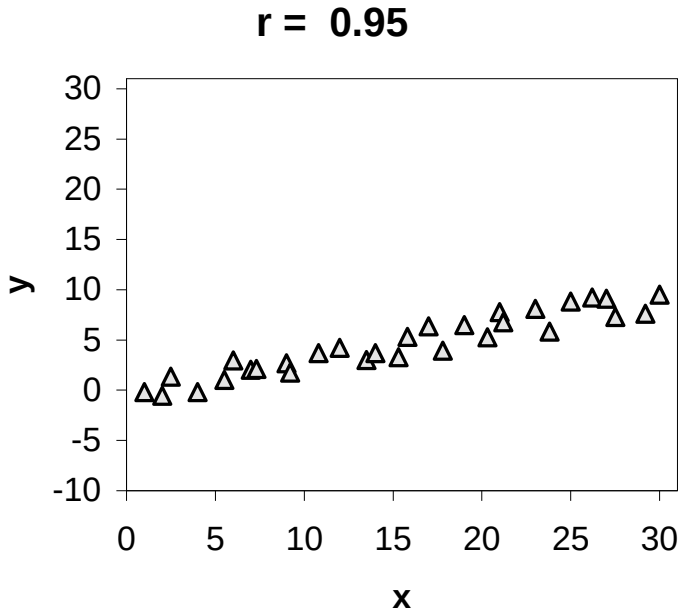
$$-1 \leq r_{xy} \leq 1$$

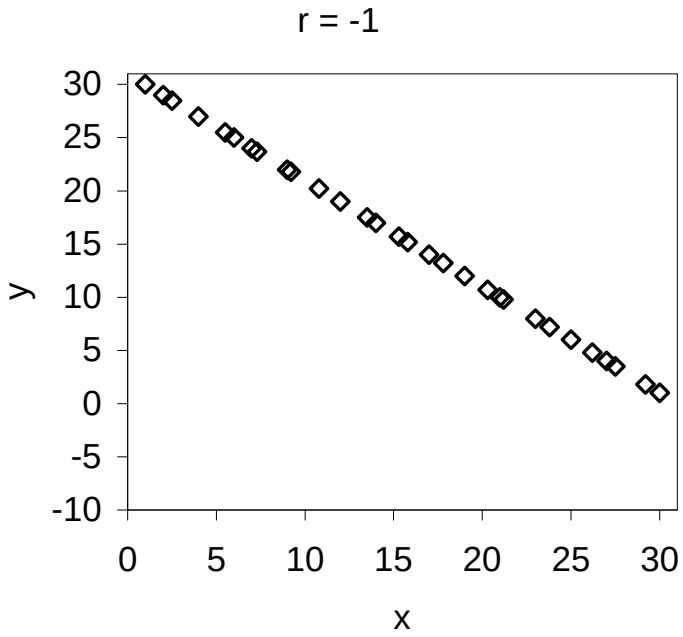
Plus il est proche de -1 ou 1, plus les données seront alignées sur une droite. Plus il est proche de 0, plus les données seront dispersées dans le plan.

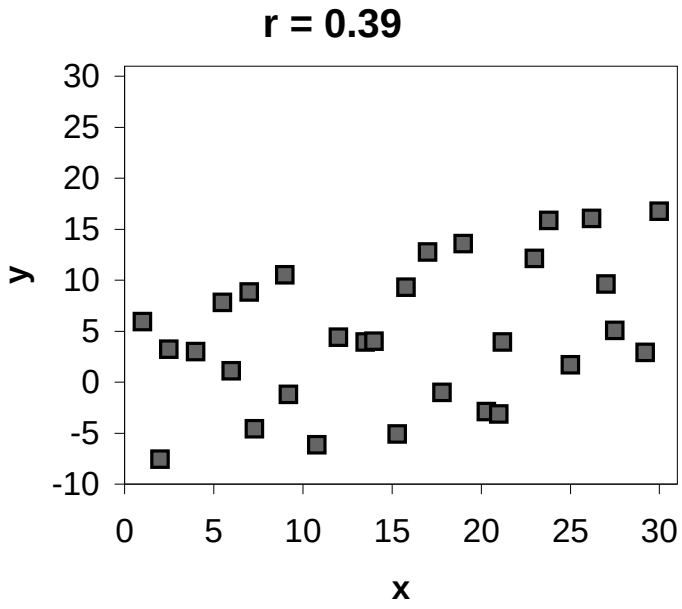
- $r_{xy} = \pm 1 \Leftrightarrow$ relation linéaire parfaite entre les deux variables : droite $y(x) = b_0 + b_1 \cdot x$
- $r_{xy} > 0 \Leftrightarrow$ les deux variables évoluent dans le même sens (si x augmente, y augmente)
- $r_{xy} < 0 \Leftrightarrow$ les deux variables évoluent en sens contraire (si x augmente, y diminue, et inversement)

Illustrations

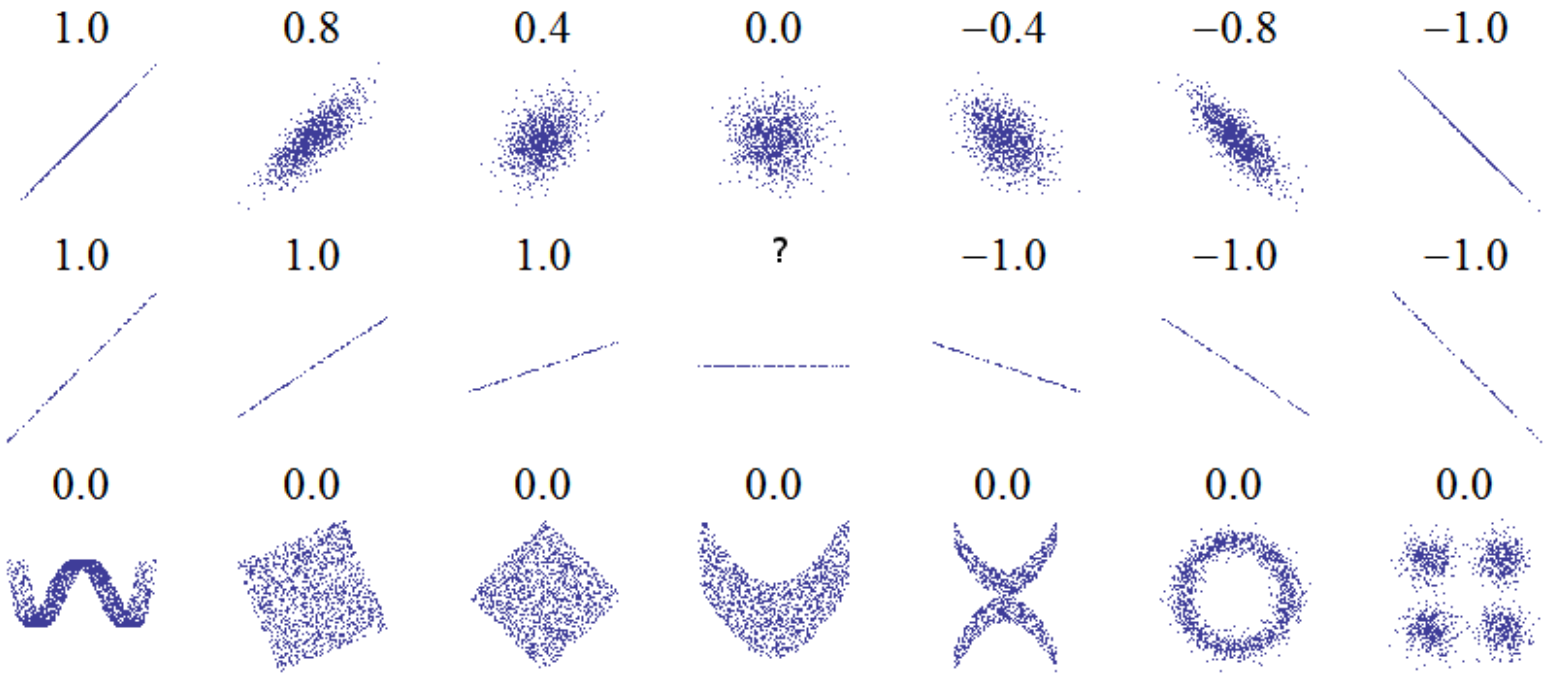








Illustrations



Remarque : la valeur « ? » est due au fait que $y_i = \bar{y}$ pour tout i (donc $s_y = 0$). Toutefois la dépendance est clairement linéaire.

Moralité : toujours faire la représentation graphique (nuage de point) des variables conjointes, pour se faire une idée a priori de la situation.

Validité d'une corrélation

Test a priori sur l'existence ou non d'une corrélation linéaire :

$$\begin{aligned} H_0 &: \rho = \rho_0 = 0 \\ H_1 &: \rho = \rho_1 \neq 0 \end{aligned}$$

Si l'on souhaite tester une corrélation positive ou négative, on fera respectivement un test unilatéral à droite ou à gauche.

Le test à considérer est celui de Student. La statistique de test est :

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad dl = n - 2$$

avec : r = coefficient de corrélation linéaire
n = taille de l'échantillon

Bien entendu, la p-valeur permet aussi de conclure.

Régression linéaire simple

Le *modèle de régression linéaire simple* est défini par l'équation suivante :

$$y = \beta_0 + \beta_1 x + \epsilon$$

avec

y	=	variable dépendante (ou variable expliquée), $y = y(x)$
x	=	variable indépendante (ou variable explicative)
β_0	=	constante de la droite de régression pour la population (ordonnée à l'origine)
β_1	=	pente de la droite de régression pour la population
ϵ	=	terme d'erreur (ou résidu)

Hypothèses :

- les erreurs sont i.i.d. selon une loi normale $\mathcal{N}(0, \sigma^2)$
- la relation linéaire entre les deux variables est légitime

Estimation du modèle

Méthode des moindres carrés : la somme des carrés des résidus (distance verticale à la droite) est minimisée.

Animation interactive.

La droite d'ajustement s'écrit :

$$\hat{y} = b_0 + b_1 x$$

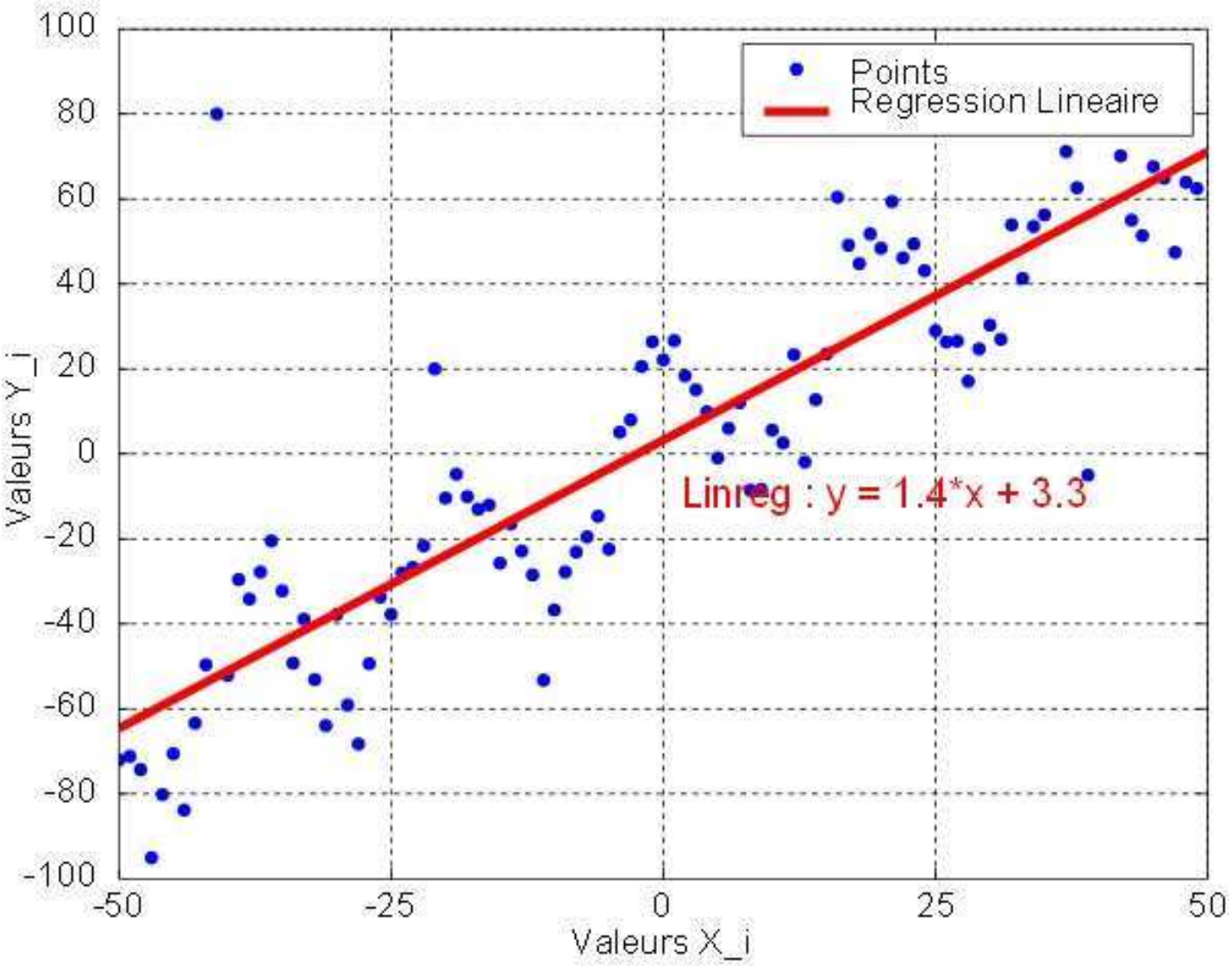
avec : $\hat{y}$ = valeur estimée de y
x = valeur de la variable indépendante

b_0 et b_1 sont des estimateurs sans biais des paramètres β_0 et β_1 et se calculent ainsi :

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad \text{ou :} \quad b_1 = r_{xy} \cdot \frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{(la droite passe par le point milieu } (\bar{x}, \bar{y})\text{)}$$

Illustration



Adéquation du modèle

Un outil pour vérifier l'adéquation du modèle est le coefficient de détermination, noté R^2 : c'est la proportion de variation totale dans la variable dépendante y qui est expliquée par sa relation avec la (ou les) variable(s) indépendante(s) x.

Dans le cas d'une seule variable indépendante, le coefficient de détermination est simplement égal au carré de la valeur du coefficient de corrélation linéaire de Pearson :

$$R^2 = r_{xy}^2$$

C'est une mesure de la qualité d'un modèle de régression linéaire. Et puisque r_{xy}^2 varie entre -1 et $+1$, R^2 varie entre 0 et 1. Plus R^2 est proche de 1, meilleur est le modèle. En pratique, des valeurs supérieures ou égales à 0.7 indiquent que le modèle est satisfaisant.

Ajustement de R^2

Lorsque la taille de l'échantillon n est petite, la valeur de R^2 est surestimée : il est préférable d'utiliser le coefficient de détermination ajusté R^2_{adj}. Pour des données bivariées, ce coefficient se calcule comme

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-2}$$

Si on vous le dit...

UTILISEZ LE R^2 AJUSTÉ !

MARRE DU R^2 ? Comme monsieur Statos, optez pour une qualité de régression plus sûre !!!

« Avant, j'utilisais un R^2 normal, j'étais fatigué et ça se voyait sur mon visage ; depuis que j'ai découvert le R^2 ajusté, ma vie a complètement changé ! »

Dépêchez-vous !

SATISFAIT ou REMBOURSÉ (*)

VU SUR INTERNET !!!

Dernière minute :

Pour vous souhaiter la bienvenue, la somme des carrés des résidus vous est offerte !

(*) voir conditions au verso

Exemple : la Tour de Pise

La Tour de Pise ne cessait de s'incliner avant d'être stabilisée. Existait-il une relation linéaire entre l'inclinaison de la Tour et l'année, avant les travaux de stabilisation ?

Exemple (suite)

L'inclinaison de la Tour a été relevée de 1975 à 1987 :

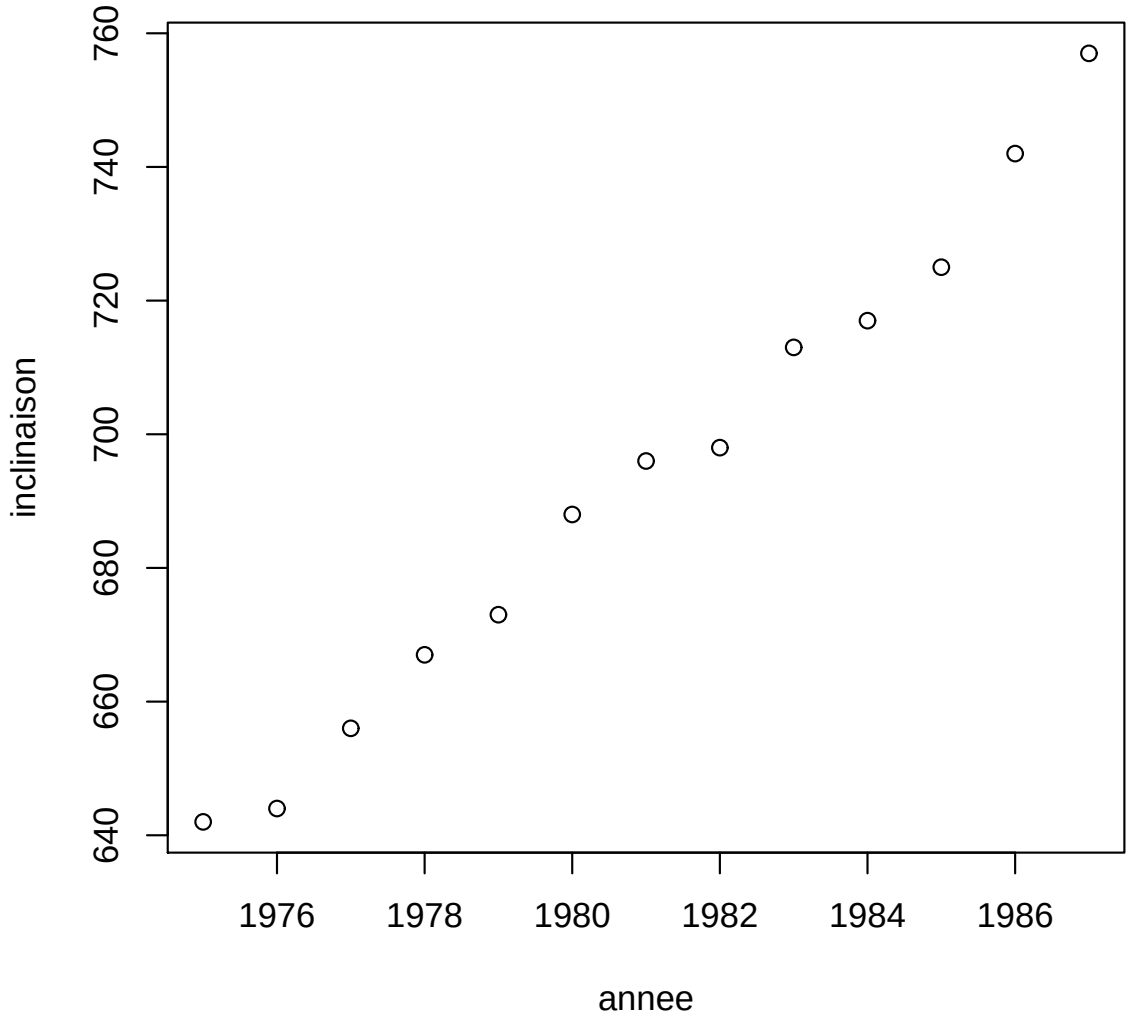
	année	inclinaison
1	1975	642
2	1976	644
3	1977	656
4	1978	667
5	1979	673
6	1980	688
7	1981	696
8	1982	698
9	1983	713
10	1984	717
11	1985	725
12	1986	742
13	1987	757

Exemple (suite)

- La première valeur pour la variable inclinaison est 642. Elle correspond à 2.9642 m et il s'agit en fait de la distance entre un point de référence si la Tour de Pise était droite et le point correspondant de la Tour à l'année considérée.
- L'inclinaison est la variable de réponse y (variable dépendante), et l'année est la variable explicative x (variable indépendante).
- Question : peut-on prédire l'inclinaison de la Tour à partir de l'année ?

Exemple (suite)

Les données peuvent être représentée par un nuage de points, ce qui nous donne une première impression visuelle :



Exemple (suite)

Les données semblent bien s'ajuster à un modèle linéaire. Notons A les années, et I les inclinaisons.

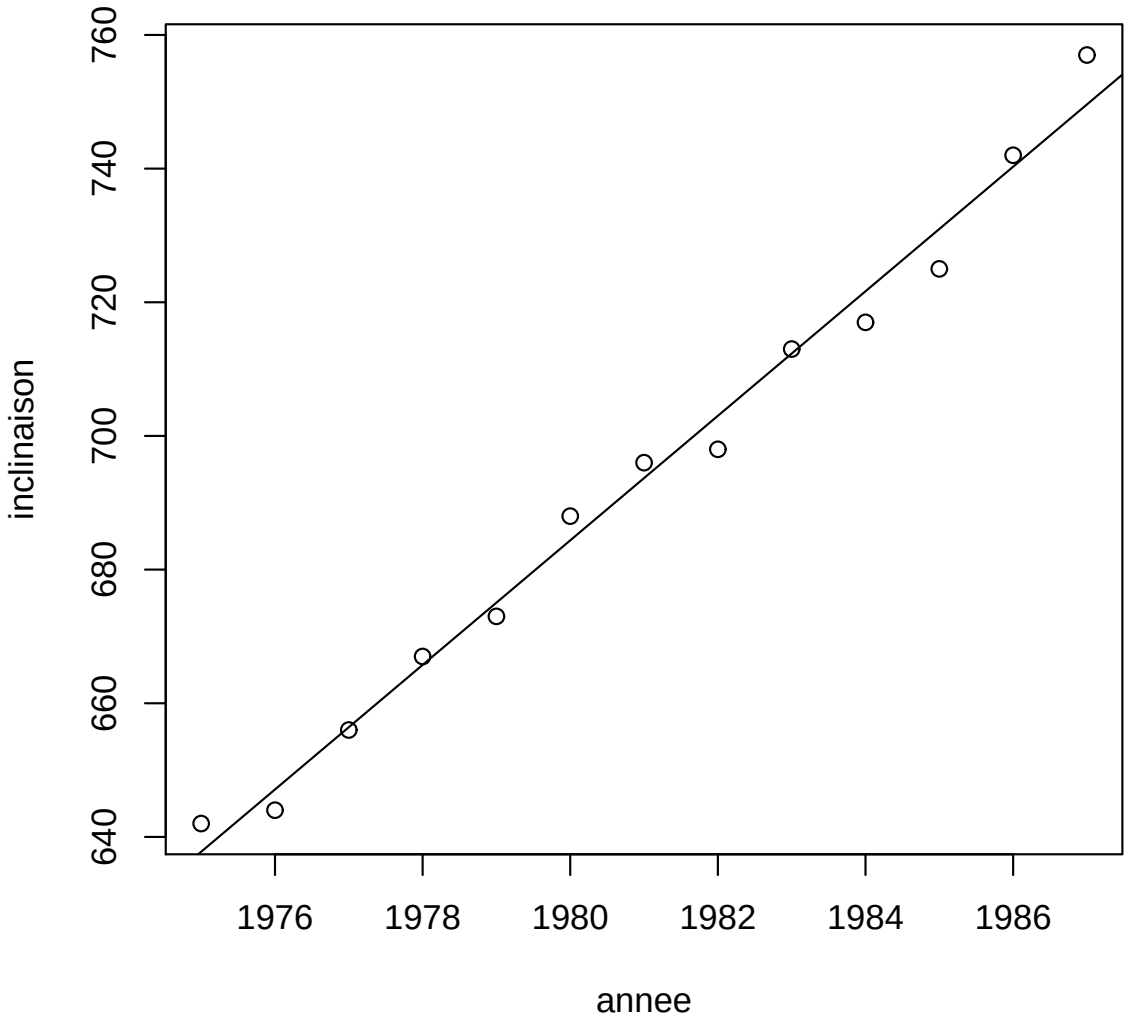
Droite de régression et coefficient de détermination :

moyennes :	$\bar{A} = 1981,\ \bar{I} = 693.692$
écarts-type :	$s_A = 3.894,\ s_I = 36.511$
covariance :	$s_{AI} = 141.333$
coeff. de corrélation :	$r_{AI} = 0.994$
droite de régression :	$y = -17766.615 + 9.319 \cdot x$
coeff. de détermination :	$R^2 = 0.9880$
R^2 ajusté :	$R^2_{adj} = 0.9869$

Test sur la corrélation : $dl = 11$, statistique $t = 30.06858$ (plus grand que la valeur critique pour n'importe quel α) et une p-valeur de $6.50 \cdot 10^{-12} \simeq 0$.

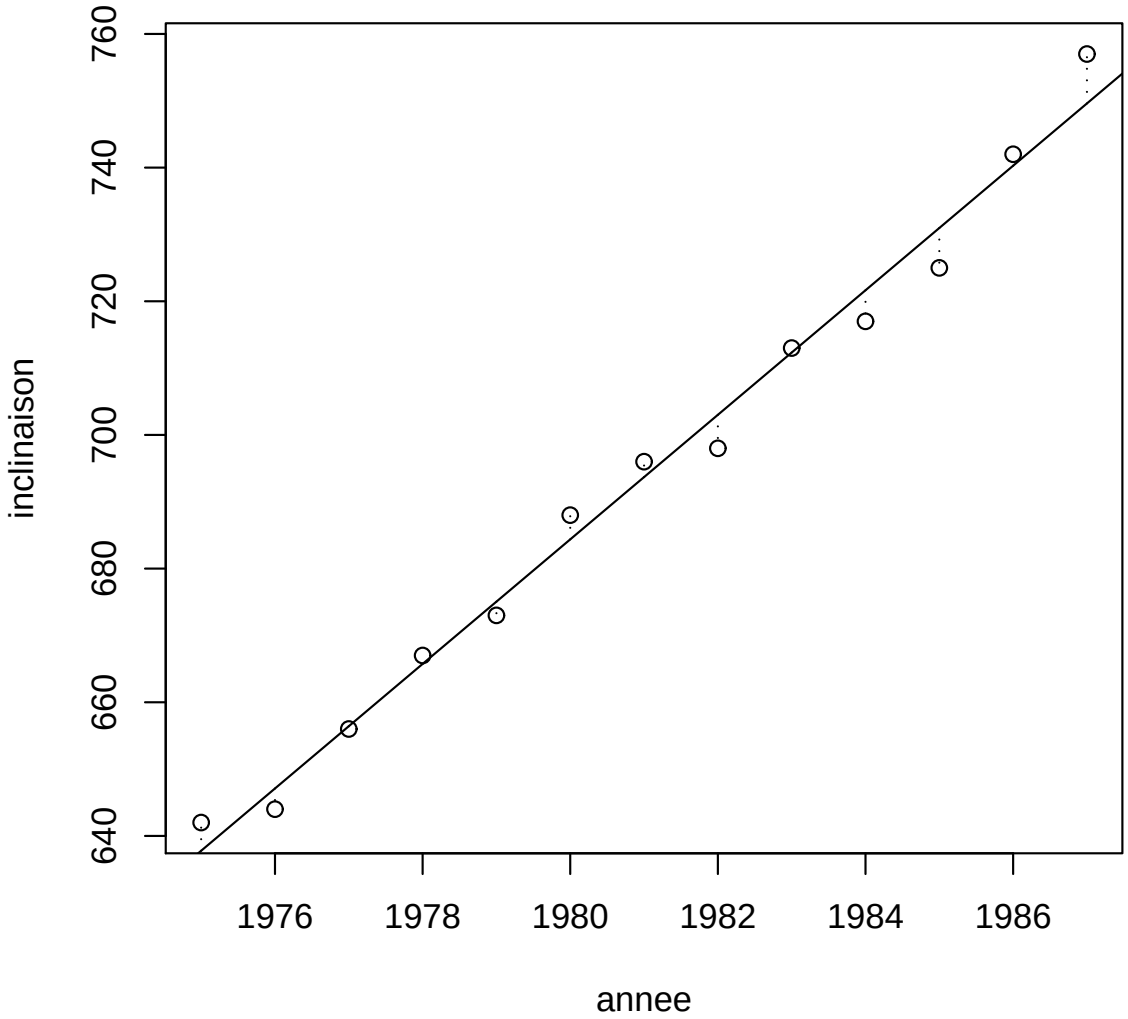
Exemple (suite)

Droite de régression : $y = -17766.615 + 9.318681 \cdot x$



Exemple (suite)

Cette droite minimise la somme des carrés des résidus :



Exemple (suite)

Le coefficient de détermination (ajusté ou non) vaut environ 0.99, ce qui est excellent.

En particulier, cela indique que près de 99% de la variabilité de l'inclinaison est expliquée par (la régression de l'inclinaison de la Tour sur) l'année.

Ainsi, bien que la régression soit en mesure d'expliquer une grande partie de la variabilité de l'inclinaison de la Tour de Pise, elle ne parvient pas à le faire totalement. D'autres facteurs peuvent intervenir.

Résultat logiciel R


```
Call:
lm(formula = inclinaison ~ annee)

Residuals:
     Min       1Q   Median       3Q      Max
 -5.9670  -3.0989   0.6703   2.3077   7.3956

Coefficients:
                Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)  -1.777e+04    6.139e+02    -28.94   9.86e-12 ***
 annee         9.319e+00    3.099e-01     30.07   6.50e-12 ***
 ---
 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.181 on 11 degrees of freedom Multiple
R-squared: 0.988,   Adjusted R-squared: 0.9869 F-statistic: 904.1
on 1 and 11 DF,   p-value: 6.503e-12
```

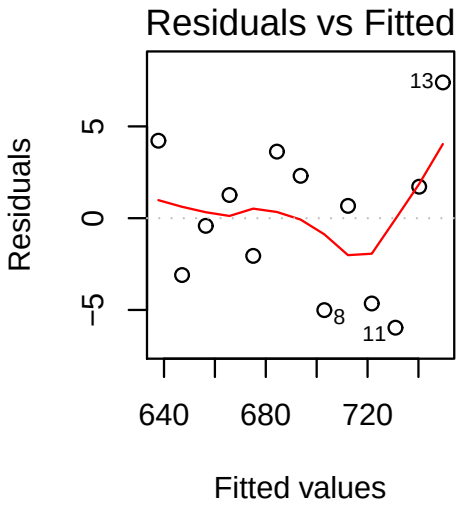
Validation des hypothèses du modèle

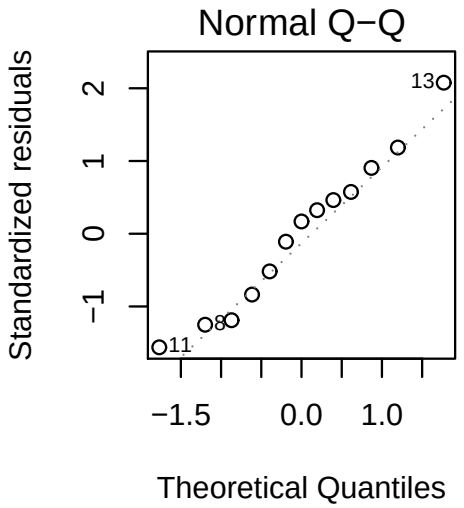
Nous avons déjà mentionné les hypothèses inhérentes au modèle de régression linéaire :

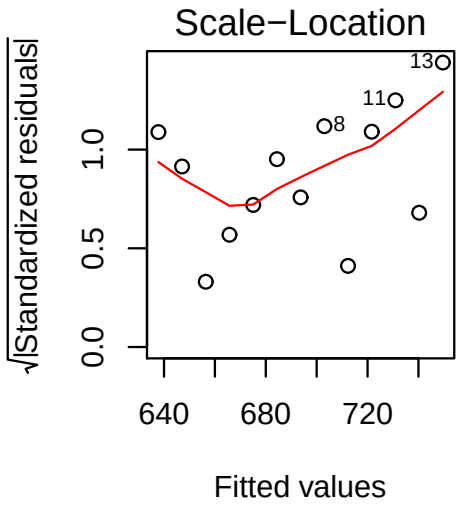
- Linéarité de la relation.
- Nullité de l'espérance des erreurs ε_i et leur variance constante σ^2.
- Normalité des variables aléatoires erreurs ε_i.

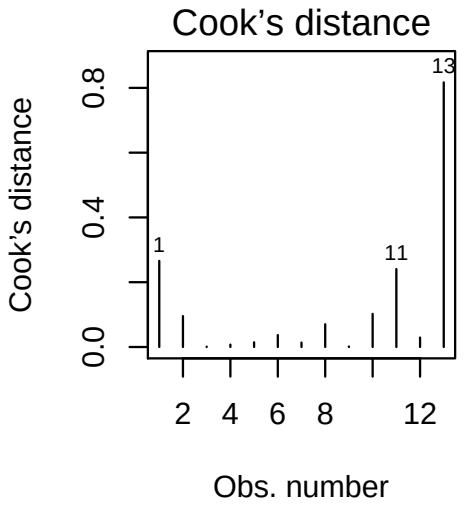
Quatre graphiques permettent de vérifier ces hypothèses. Leur construction est essentielle.

Validation des hypothèses (suite)









Validation des hypothèses (suite)

- Graphiques 1. et 3. : résidus contre valeurs ajustées.

  $\rightsquigarrow$ vérification de l'ajustement du modèle (si une tendance se dessine, le modèle n'est pas adéquat) ; vérification des hypothèses sur l'espérance nulle et la variance constante (une variabilité semblable doit se dessiner autour de 0 dans le graphique 1.), et détection de valeurs atypiques (indiquées par leur numéro d'observation)

- Graphique 2. : résidus standardisés contre résidus théoriques.

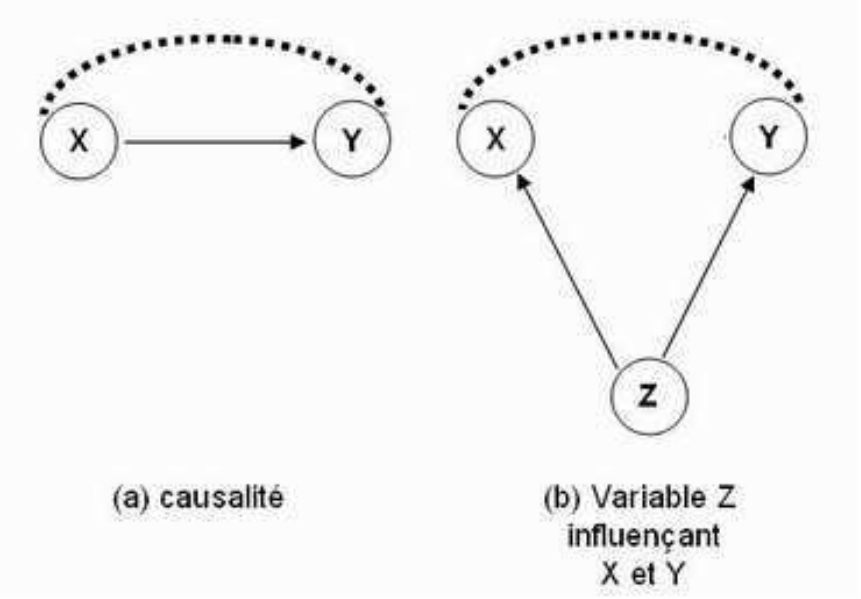
  $\rightsquigarrow$ vérification de la normalité des erreurs (si l'hypothèse est vraie, les points ne doivent pas s'éloigner d'une droite) et détection de valeurs atypiques

- Graphique 4. : distance de Cook.

  $\rightsquigarrow$ identification de possibles valeurs atypiques

Pièges de la régression linéaire

Une corrélation forte n'implique pas nécessairement causalité. En effet, les deux variables peuvent être influencées par une troisième variable. Elle pousse davantage à la réflexion et à de nouvelles investigations.



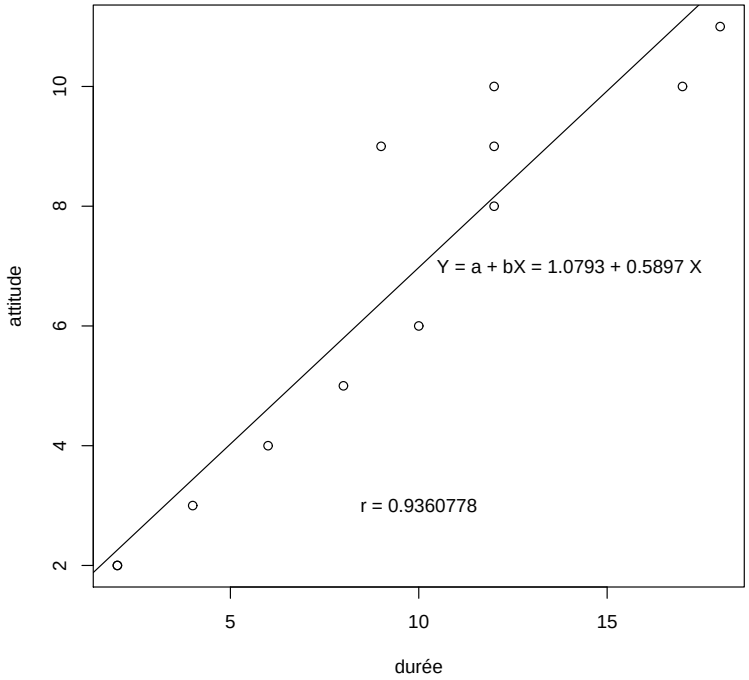
À l'inverse, une corrélation nulle indique uniquement qu'il n'y a pas de relation *linéaire*. D'autres relations peuvent exister.

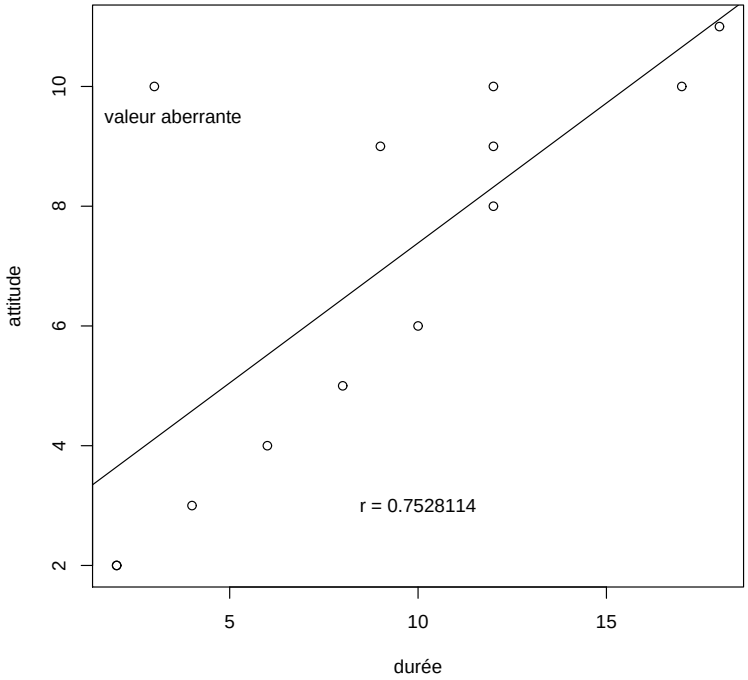
Exemples

- Pour l'ensemble des communes d'Alsace, il a été observé une étonnante corrélation entre le nombre de naissances et celui des cigognes recensées sur les cheminées. Est-ce à dire que les enfants alsaciens ont été apportés par les cigognes ?
- Les services de santé ont observé une corrélation positive entre le taux d'utilisation de crème solaire et le risque de cancer de la peau. Les crèmes solaires seraient-elles cancérigènes ?
- La parabole $y = x^2$ implique une corrélation nulle, mais les deux variables sont complètement dépendantes !

Valeurs atypiques

Attention aux valeurs atypiques : elles influencent fortement la droite de régression et nécessitent une étude particulière !





La constatation visuelle parle d'elle-même. Selon le critère du coefficient de détermination, dans le premier cas le modèle est assez bon, avec un $R^2 = 0.8762$, tandis que le deuxième est inutilisable avec un $R^2 = 0.5667$

Régression multiple

Le modèle de régression peut être étendu à plusieurs variables explicatives x_i. On parle alors de régression linéaire multiple.

- En 1984, le temps record de plusieurs courses de montagnes en Écosse a été relevé. On se demandait alors s'il existait une relation linéaire entre le temps record, la longueur et la dénivellation totale du parcours.
- Une enquête a été menée pour prédire la consommation des véhicules, exprimée en MPG (miles parcourus par gallon de carburant, plus le chiffre est élevé, moins la voiture consomme), à partir de leurs caractéristiques : poids, rapport de pont, puissance, etc. Au final, les variables significatives étaient le poids (weight) et le rapport de pont (drive ratio). Les autres semblaient sans effet dans l'explication de la consommation.

  Cette lecture très simplifiée du rôle des variables doit bien sûr être relativisée (cf. pièges). La puissance (horsepower) est vraisemblablement masquée par le poids auquel elle est très fortement corrélée.

Analyse d'une régression linéaire simple

Les étapes suivantes montrent comment effectuer une analyse de régression linéaire simple :

1. Définir la variable dépendante y et la variable indépendante x
2. Dessiner un diagramme de dispersion sur x et y afin de vérifier visuellement si une relation linéaire est probable.
3. Calculer le coefficient de corrélation r_{xy} et le coefficient de détermination R^2.
4. Effectuer un test statistique pour valider ou invalider le modèle linéaire simple. Commenter les graphiques s'ils sont à disposition.
5. Calculer la droite de régression.
6. Établir une conclusion.