h e g

Haute école de gestion de Genève
Geneva School of Business Administration

Régression linéaire simple

Dr Sacha Varone

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).

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

Rejet de H_0 = distribution théorique
si une trop grande différence existe.

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.

Test d'indépendance

But : Tester l'indépendance de deux variables de type catégorielles.
Moyen : Table de contingence.

Principe du test :
Comparaison des fréquences observées avec les fréquences théoriques e_i en cas d'indépendance.
Si une trop grande différence existe, alors l'hypothèse d'indépendance des variables est rejetée.

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

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, effectif observé par cellule ≥ 5.

Régression linéaire

Modèle 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)
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 ;
- la relation linéaire entre les deux variables est légitime.

Estimation du modèle

Méthode des moindres carrés : la somme des carrés des résidus est minimisée.

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

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Coefficient de corrélation

Le *coefficient de corrélation linéaire* de Pearson d'un échantillon est la valeur

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

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

Une corrélation linéaire r étant le plus souvent calculée à partir d'un échantillon, 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.

Remarque. La dernière formule permet un calcul plus rapide.

Validité d'une corrélation

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

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

La statistique de test à considérer est la suivante :

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

avec

t = nombre d'écart-type de r depuis 0
r = coefficient de corrélation linéaire
n = taille de l'échantillon

Exemple

Une entreprise souhaite analyser la relation entre la taille d'une annonce publicitaire, et le nombre d'appels reçus générés par l'annonce. Elle veut savoir s'il existe une corrélation linéaire positive entre ces deux variables, au seuil 0.05. Pour cela, elle demande à ses clients d'indiquer quelle annonce leur a fait connaître l'entreprise.

Le dépouillement de l'enquête donne :

Taille [cm^2]	90	160	250	160	200	...
Prop. appels	0.13	0.16	0.21	0.18	0.18	...

Taille [cm^2]	...	160	200	200	160	90
Prop. appels	...	0.19	0.15	0.17	0.13	0.11

Exemple (suite)

1. Le paramètre d'intérêt est la corrélation linéaire ρ entre la taille d'une annonce publicitaire et la proportion d'appels générés par l'annonce.
2. L'hypothèse nulle et alternative sont :

$$\begin{aligned} H_0 &: \rho \leq 0 \\ H_1 &: \rho > 0 \end{aligned}$$

3. Le niveau de signification choisi est $\alpha = 0.05$
4. La p-valeur associée est 0.003921
5. Comme la p-valeur est inférieure au niveau de signification, l'hypothèse nulle est rejetée.
6. L'échantillon suppose la possibilité d'une relation linéaire positive entre la taille d'une annonce publicitaire et la proportion d'appels générés par l'annonce.

Qualité d'un modèle linéaire

On définit :

- La somme des carrés totale (SST = *Total Sum of Squares*) :

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- La somme des carrés des erreurs (SSE = *Sum of Squares Errors*) :

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

- La somme des carrés de régression (SSR = *Sum of Squares Regression*) :

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

où

n = taille de l'échantillon
y_i = i-ème valeur de la variable dépendante
$\hat{y}_i$ = i-ème valeur prédite
$\bar{y}$ = moyenne de la variable dépendante

Variance totale

La variance totale se décompose en une partie expliquée et une partie non-expliquée (ou résiduelle) :

$$\text{SST} = \text{SSE} + \text{SSR}$$

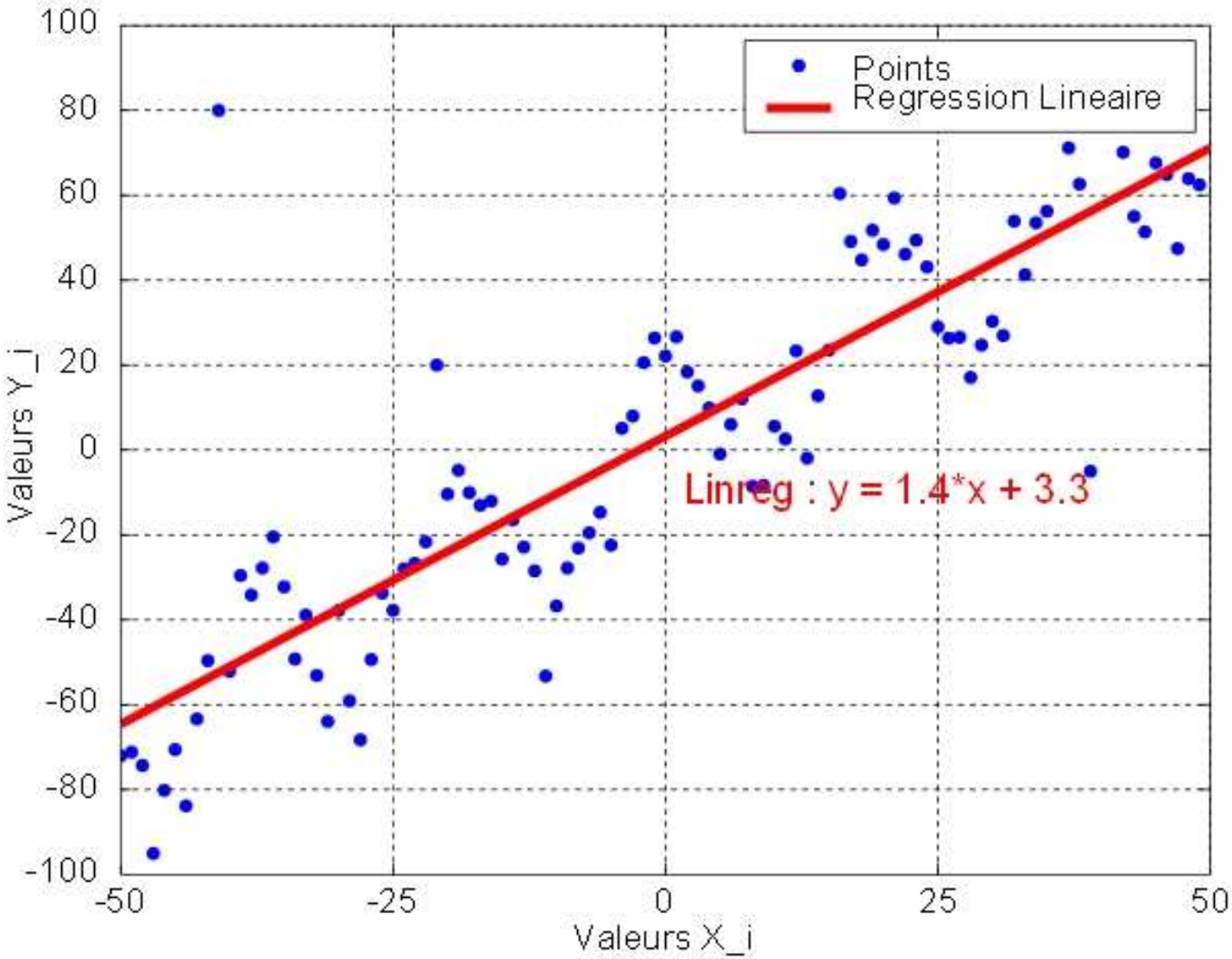
Cette propriété est à la base de l'analyse de variance, utilisée pour tester si plusieurs populations sont significativement différentes les unes des autres.

Propriété

Soit une droite de régression linéaire, basée sur la minimisation de la somme des carrés.

- La somme des résidus est nulle :
$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$
- La somme des carrés des résidus (SSE) est minimale.
$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
- La droite de régression linéaire simple passe par le point $(\bar{x}, \bar{y})$.
- Les coefficients estimés b_0 et b_1 sont des estimateurs sans biais de β_0 et β_1.

Illustration



Coefficient de détermination

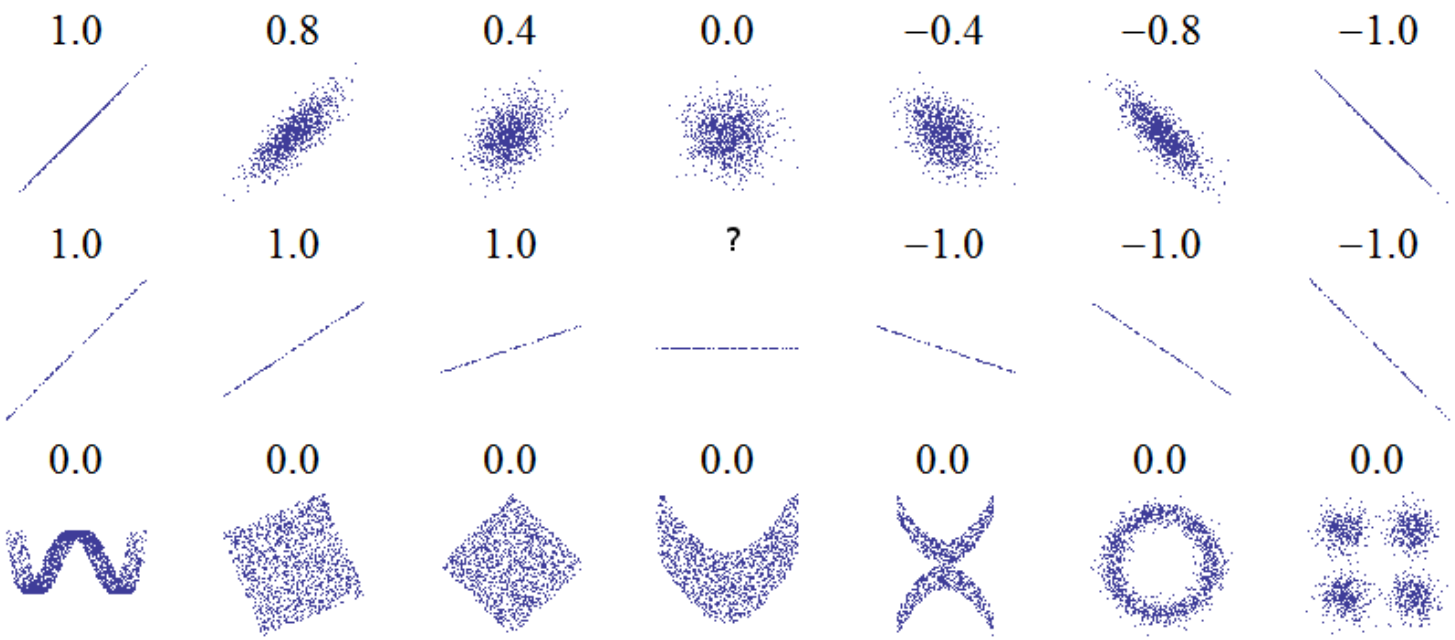
Le *coefficient de détermination* R^2 est la proportion de variation totale dans la variable dépendante qui est expliquée par sa relation avec la variable dépendante.

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

C'est une mesure de la qualité d'un modèle de régression linéaire. R^2 varie entre 0 et 1. Plus il se rapproche de 1, meilleur est le modèle. En pratique, des valeurs supérieures ou égales à 0.7 indiquent que le modèle est satisfaisant.

Remarque : dans le cas d'une seule variable indépendante, le coefficient de détermination est égal au carré de la valeur du coefficient de corrélation linéaire de Pearson : $R^2 = r_{xy}^2$

Illustration



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

Exemple (suite)

L'entreprise souhaitant analyser la relation entre la taille d'une annonce publicitaire et le nombre d'appels reçus générés par l'annonce est parvenue à la conclusion qu'une relation linéaire existe probablement.

Le coefficient de détermination vaut

$$R^2 = r_{xy}^2 = 0.7795506$$

Ainsi, environ 78% de la variance du nombre d'appels est expliquée par la taille de l'annonce.

Validité du modèle

Pour tout modèle, il est nécessaire d'en vérifier sa validité. Autrement dit, il est nécessaire de connaître si le modèle est statistiquement significatif. Pour le modèle linéaire simple, en supposant valides les postulats sur les résidus (i.i.d. et suivant une loi normale), il existe deux méthode de test équivalentes :

1. Test de signification de la corrélation entre x et y (déjà vu)
2. Test de signification du coefficient de la pente de régression

Test sur le coefficient de la pente de régression

On utilise les hypothèses nulle et alternative suivantes :

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

L'estimation de l'écart type de la pente de régression est

$$s_{b_1} = \frac{s_\epsilon}{\sqrt{(x - \bar{x})^2}} \qquad \text{avec} \quad s_\epsilon = \sqrt{\frac{SSE}{n-2}}$$

La variable de test à utiliser est

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \quad dl = n - 2$$

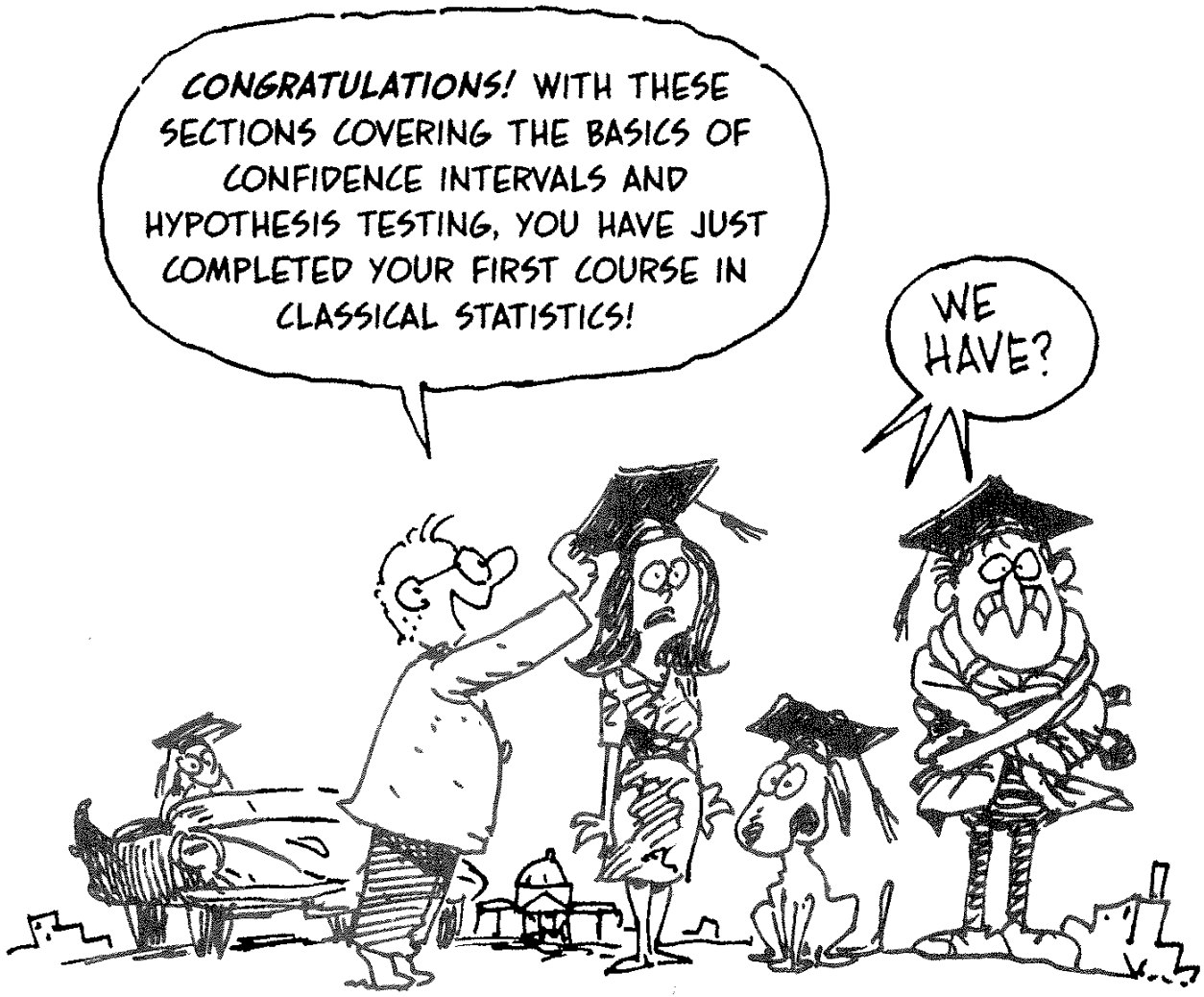
avec

β_1 = pente supposée de la droite de régression
b_1 = pente calculée de la droite de régression
s_{b_1} = estimation de l'écart type de la pente de régression

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}
4. Calculer les coefficients de la régression, ainsi que le coefficient de détermination R^2
5. Effectuer un test statistique pour valider ou invalider le modèle linéaire simple.
6. Établir une conclusion.



source : "The Cartoon Guide to Statistics", L. Gonick & W. Smith