



Régression linéaire à une variable explicative

Modèle

Thierry Dhorne

¹Institut Universitaire de Technologie de Vannes

²Université de Bretagne-Sud

Licence - Mise à niveau

Année Universitaire 2014-2015



Allure générale

Régression à une variable

Modèle

Principes

Modèle

Linéarité

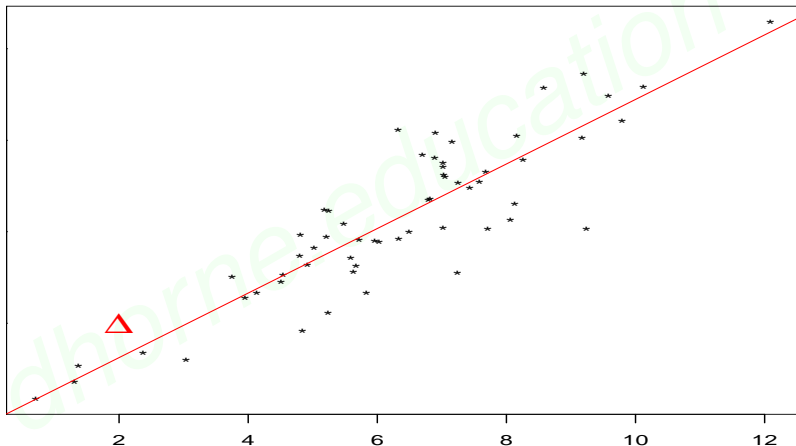
Postulats

Non biais

Variance constante

Indépendance

- on observe une allure générale linéaire (affine)
- ➔ que l'on modélise par une droite Δ





Écarts spécifiques

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

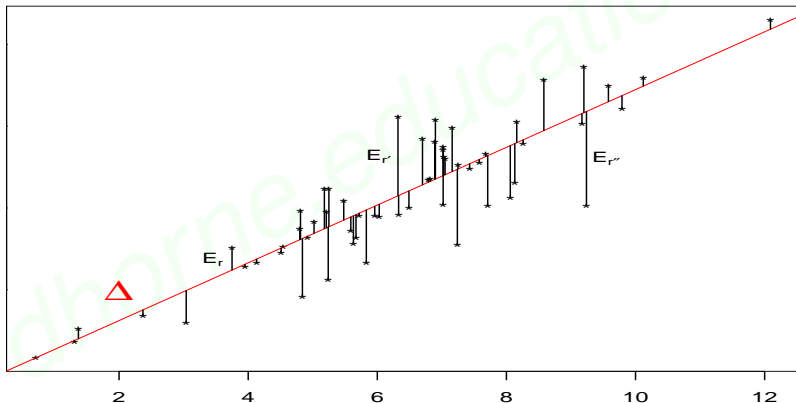
Postulats

Non biais

Variance
constante

Indépendance

- on voit que chaque observation s'écarte de la droite de façon spécifique (traits verticaux)
- ➔ on modélise l'écart par une variable aléatoire qui formalise cette fluctuation





- le modèle utilisé s'écrit

Modèle de régression linéaire simple

$$Y_r = \beta_0 + \beta_1 x_r + E_r, \quad r = 1, \dots, n$$

- ▶ Y est la variable à expliquer
- ▶ x est la variable prédictrice
- ▶ β_0 et β_1 représentent les deux paramètres du modèle
- ▶ les E_r représentent les écarts au modèle et sont considérées comme des variables aléatoires
- ▶ n est la taille de l'échantillon ou nombre d'observations



Modèle et interprétation

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

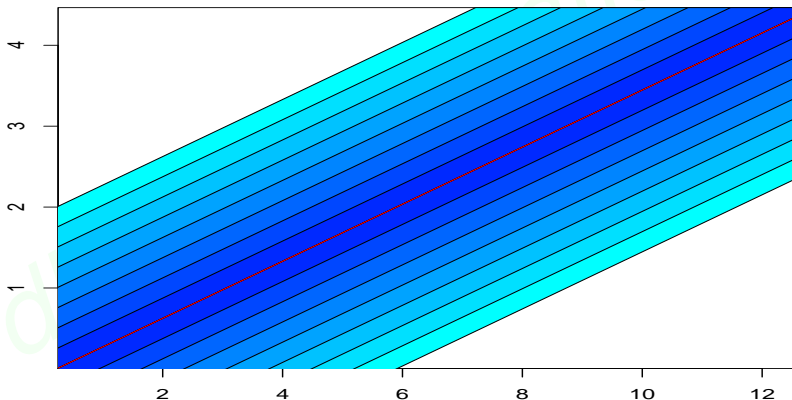
Postulats

Non biais

Variance
constante

Indépendance

- le modèle peut être représenté par
 - ▶ une espérance linéaire (affine)
 - ▶ une dispersion aléatoire autour de cette espérance (dont la probabilité est d'autant plus faible que l'on s'éloigne de l'axe)





Lien linéaire

Régression à une variable

Modèle

Principes

Modèle

Linéarité

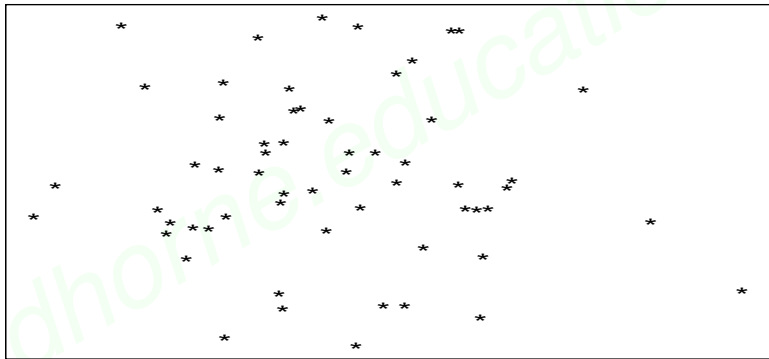
Postulats

Non biais

Variance constante

Indépendance

- l'aspect central du modèle est le lien linéaire à l'instar de ce qui se passe pour les moules
- ★ parfois il peut ne pas exister de lien entre deux variables comme sur le graphique suivant



→ la connaissance de x n'apporte aucune information sur Y



- le modèle qui vient d'être présenté

$$Y_r = \beta_0 + \beta_1 x_r + E_r$$

s'accompagne de trois postulats

Postulats du modèle de régression linéaire simple

Les E_r sont des variables aléatoires telles que

- ▶ $E(E_r) = 0 \Rightarrow E(Y_r) = \beta_0 + \beta_1 x_r$
- ▶ $V(E_r) = V(Y_r) = \sigma_E^2$
- ▶ $C(E_r, E_{r' \neq ir}) = 0$



Postulat de non biais (d'espérance linéaire)

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

Postulats

Non biais

Variance
constante

Indépendance

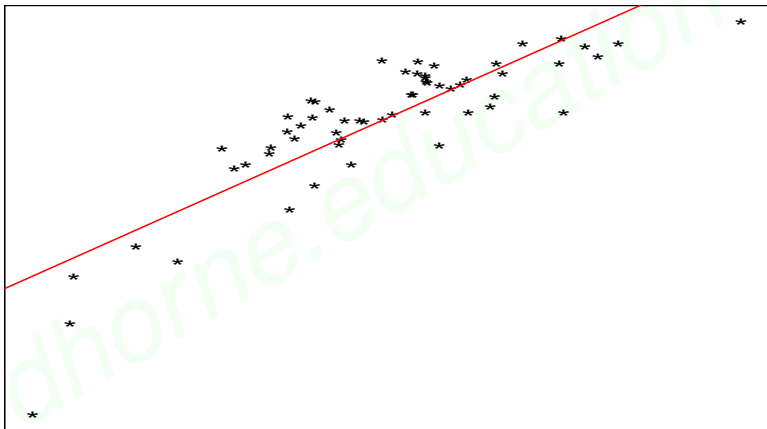
- le postulat $E(Y_r) = \beta_0 + \beta_1 x_r$,
- indique qu'il n'y a pas de biais dans le modèle
- et ne paraît donc pas très restrictif en soi
- ★ mais attention car on peut avoir oublié d'autres variables
- comme on le verra plus tard



Lien non linéaire

Non détecté

- de la même manière il peut y avoir un biais dans le modèle si le lien n'est pas linéaire



Régression à
une variable

Modèle

Principes

Modèle

Linéarité

Postulats

Non biais

Variance
constante

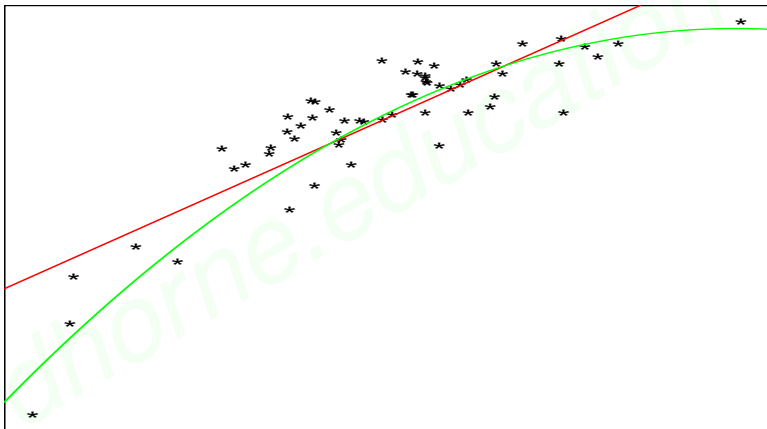
Indépendance



Lien non linéaire

DéTECTÉ

- de la même manière il peut y avoir un biais dans le modèle si le lien n'est pas linéaire



Régression à
une variable

Modèle

Principes

Modèle

Linéarité

Postulats

Non biais

Variance
constante

Indépendance



Postulat de variance constante

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

Postulats

Non biais

Variance
constante

Indépendance

- le postulat $V(Y_r) = \sigma_E^2$
 - implique que la variance résiduelle est constante (ne dépend pas de x)
 - et donc est commune à toutes les observations
 - ★ ceci est une contrainte assez forte qu'il convient de vérifier
 - ★ nous indiquerons plus loin comment valider le postulat de variance constante
 - ce postulat s'appelle souvent postulat d'homoscédasticité
 - terme d'origine grecque qui signifie isovariance ou équivariance



Comparaison de la variance résiduelle

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

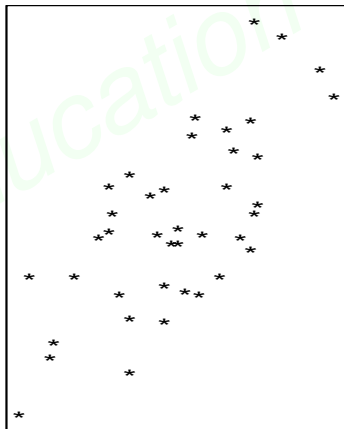
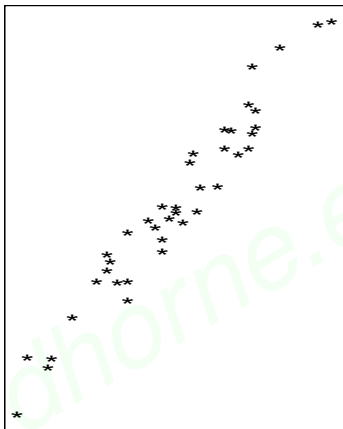
Postulats

Non biais

Variance
constante

Indépendance

- la variance doit être constante mais peut être quelconque (\pm forte)





Non homogénéité de la variance (hétéroscédasticité)

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

Postulats

Non biais

Variance
constante

Indépendance

- en revanche, il faut vérifier que la variance ne varie pas
→ par exemple avec la variable explicative





Modèle hétéroscédastique

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

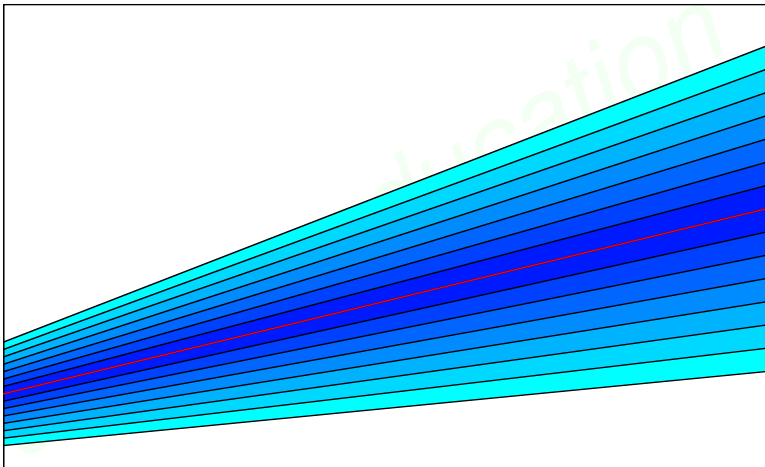
Postulats

Non biais

Variance
constante

Indépendance

- la variance n'est pas constante
- ➔ mais augmente avec l'abscisse





Postulat d'indépendance

Régression à
une variable

Modèle

Principes

Modèle

Linéarité

Postulats

Non biais

Variance
constante

Indépendance

- on postule que E_r et $E_{r' \neq r}$ sont indépendants,
 - il n'y a pas de lien probabiliste entre les observations
 - pas de structure particulière sur les données
 - en particulier pas de structure
 - ▶ temporelle (voir séries chronologiques)
 - ▶ ou spatiale (voir statistique spatiale)
- on peut obtenir l'indépendance (ou s'en approcher) par un échantillonnage adapté
 - ★ une bonne connaissance du problème suffit en général à éviter ce phénomène
 - ★ **sauf série chronologique**