Modèles linéaires, analyses factorielles et classification

Denis Laloë

Modèles linéaires

Une introduction générale

Denis Laloë
GABI - PSGen

27 septembre 2016

Introduction

Qu'est-ce qu'un modèle

- Développé dans le cadre/prolongement d'une théorie
- Origine technologique
  - Maquette
  - Représentation réduite
- Réalisation concrète, clarification, formalisation, transposition mathématique de ce qui est décrit de manière diffuse dans la science empirique
- Plus simple et plus pauvre que la réalité observée
- Signification instrumentale (outil) *Legay, 1996*

Qu'est-ce qu'un modèle

Fonction

- Organisatrice : structuration de relations entre concepts / données
- Heuristique : découverte de nouveaux faits, relations, explications
- Prévision
- Mesure

Qu'est-ce qu'un modèle

Démarche *a-modélisatrice* - Benzécri

- *Le modèle doit suivre les données, non l inverse*
- Substitution de facteurs (obtenus via une analyse factorielle des données) à *l arbitraire échafaudage des idées a priori*
- Observation vs Expérimentation
- Pas de structure a priori
- Synthèse (vision holistique)

Qu'est-ce qu'un modèle

Les mêmes caractéristiques générales

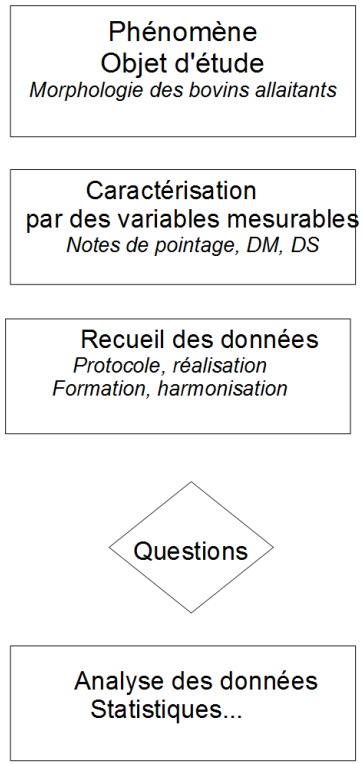
- Représentation schématique
- Fonction instrumentale
- Formalisation

Des fonctions différentes

- Cognitive
- Prédictive

- Modèle *physiologie*
  - Représentation explicative, cognitive, heuristique
- Modèle *génétique*: Modélisation statistique à partir d'un mécanisme de transmission
  - interprétation génétique
  - capacité prédictive
  - simplicité
- Modèle *statistique* : modélisation statistique
  - capacité prédictive
  - simplicité

Place des statistiques dans la caractérisation d'un phénomène



Le point de vue du statisticien : les méthodes statistiques sont faites par des gens qui n'en ont pas l'utilité, pour des gens qui n'en ont pas la maîtrise

Place des statistiques dans la caractérisation d'un phénomène

Autour des statistiques, gravitent diverses communautés, des statisticiens purs et durs aux lecteurs des résultats, avec des différences d'appréciation quant à

- l'intérêt
- l'utilité
- le jugement des méthodes et des résultats

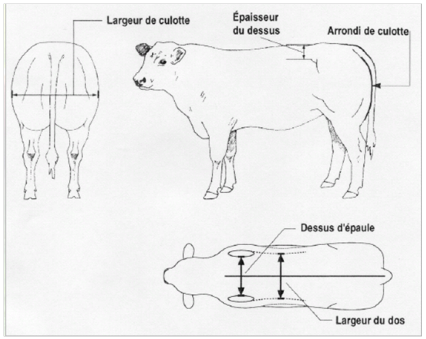
Morphologie des bovins allaitants

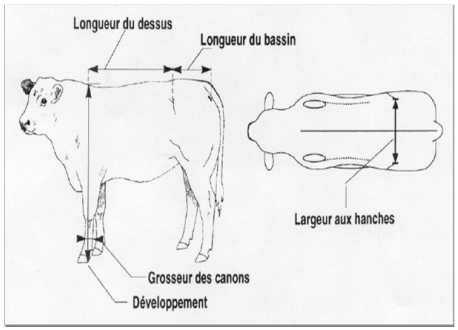
- Un animal est caractérisé par sa morphologie
- En relation avec sa "valeur"
  - bouchère (carcasse)
  - élevage (prolificité, fertilité)
  - fonctionnelle (aplombs, longévité)
  - raciale (standard)
- Important pour le choix des animaux

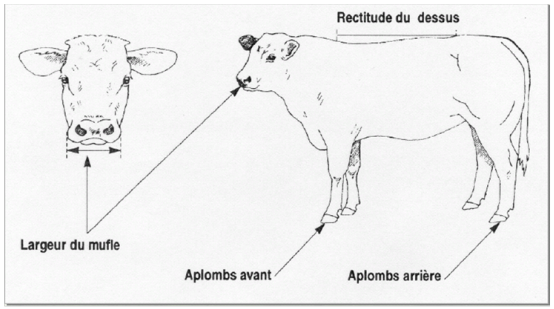
Modélisation de la morphologie

Caractérisation par des variables mesurables

- postes définis par des experts
- élevage notes données par des techniciens experts







Formation des techniciens

Formation des jeunes pointeurs

- Formation commune à toutes les races (une semaine)
- Formation spécifique raciale (une journée)
- Examen pour l'agrément
- taux de réussite à l'agrément 1/2

Harmonisation et évaluation des pointeurs confirmés

- Sessions annuelles par race
- Harmonisation
- Renouvellement d'agrément

Environ 200 000 pointages par an

Fichiers de données

- Environ 200 000 pointages par an, sur une vingtaine de postes
- Expertise : formalisation, modélisation, formation, harmonisation
- Réalisation : déplacements, salaires, traitement informatique,...
- La donnée ne tombe pas du ciel. Elle est le produit d'un investissement
  - intellectuel (expert, modélisation préalable, choix de variables, design, technologie)
  - financier

Le statisticien vs l'utilisateur

Ne pas oublier que la donnée coûte cher, et que l'utilisateur est un interlocuteur expert.

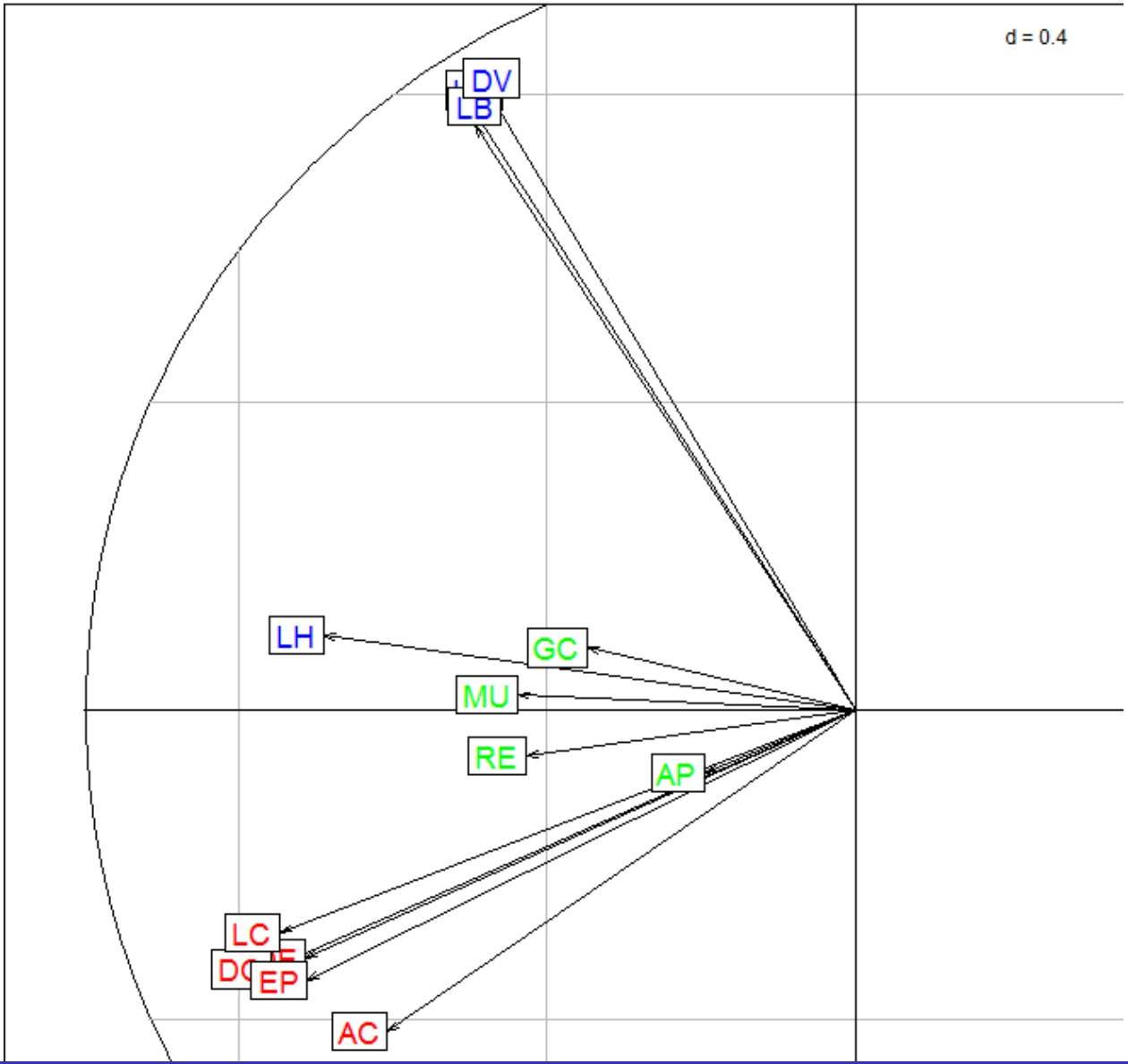
- Dialogue
- Langage commun
- Réponse à une question
- Intérêt pour l'utilisateur

Où vous situerez-vous ?

Résumé parlant ?

	DEPA	LADO	ARCU	LACU	EPDE	GRCA	LODE	LBA	LAH	DEV	TETE	APV	APR	REDE
DEPA	1	0,89	0,73	0,78	0,78	0,19	0,22	0,21	0,54	0,1	0,23	0,13	0,17	0,23
LADO	0,89	1	0,74	0,78	0,79	0,19	0,21	0,2	0,53	0,09	0,22	0,16	0,19	0,33
ARCU	0,73	0,74	1	0,81	0,8	0,06	0,09	0,07	0,43	-0,04	0,13	0,05	0,14	0,23
LACU	0,78	0,78	0,81	1	0,8	0,17	0,21	0,21	0,6	0,08	0,2	0,12	0,17	0,24
EPDE	0,78	0,79	0,8	0,8	1	0,15	0,2	0,19	0,52	0,08	0,21	0,13	0,17	0,29
GRCA	0,19	0,19	0,06	0,17	0,15	1	0,46	0,5	0,46	0,48	0,42	0,28	0,29	0,07
LODE	0,22	0,21	0,09	0,21	0,2	0,46	1	0,84	0,52	0,82	0,35	0,27	0,27	0,17
LOBA	0,21	0,2	0,07	0,21	0,19	0,5	0,84	1	0,54	0,8	0,37	0,28	0,28	0,16
LAH1	0,54	0,53	0,43	0,6	0,52	0,46	0,52	0,54	1	0,45	0,37	0,27	0,28	0,19
DEVE	0,1	0,09	-0,04	0,08	0,08	0,48	0,82	0,8	0,45	1	0,34	0,28	0,26	0,16
TETE	0,23	0,22	0,13	0,2	0,21	0,42	0,35	0,37	0,37	0,34	1	0,3	0,29	0,12
APAV	0,13	0,16	0,05	0,12	0,13	0,28	0,27	0,28	0,27	0,28	0,3	1	0,5	0,29
APAR	0,17	0,19	0,14	0,17	0,17	0,29	0,27	0,28	0,28	0,26	0,29	0,5	1	0,24
REDE	0,23	0,33	0,23	0,24	0,29	0,07	0,17	0,16	0,19	0,16	0,12	0,29	0,24	1

Résumé parlant ?



Qu'est-ce qu'un modèle linéaire

Un modèle linéaire est un modèle **statistique**

Mise en relation quantifiée de deux ensembles de variables, l'un *expliquant* l'autre

- Capacité prédictive
- Inférence (induction)
- Interprétation
- Simplicité
- Décision
- Opérationnel

Qu'est-ce qu'un modèle linéaire

Un modèle linéaire est un modèle **statistique**

Mise en relation quantifiée de deux ensembles de variables, l'un *expliquant* l'autre

- Capacité prédictive
- ...

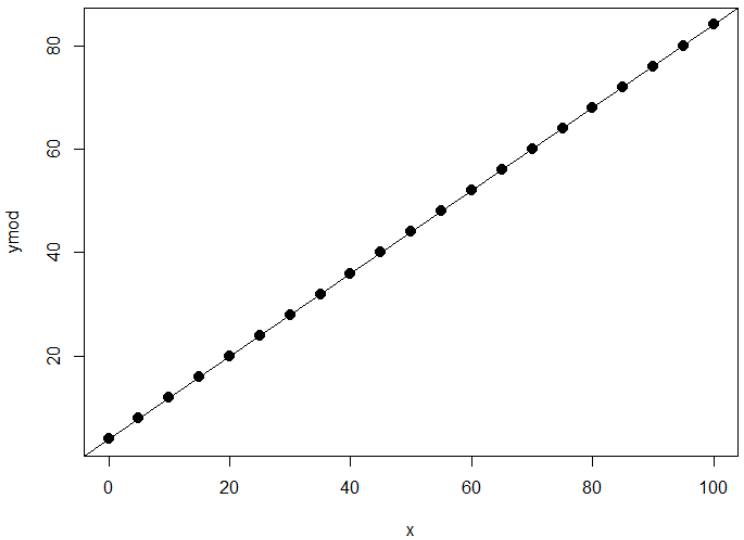
Mise en relation quantifiée de deux ensembles de variables, l'un *expliquant* l'autre

Il existe un modèle expliquant/résumant **Y** par **X**
Y est une combinaison linéaire des **X**, plus une *erreur/résidu/résiduelle*

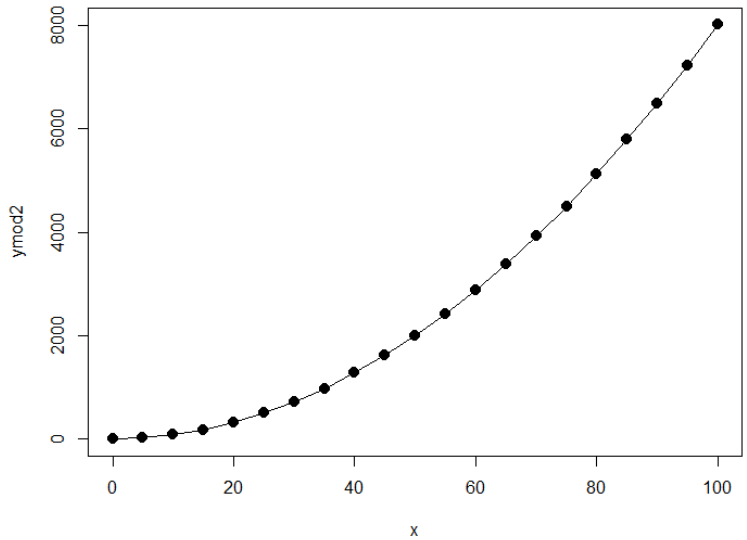
Qu'est-ce qu'un modèle linéaire

On modèlise une variable y par une fonction linéaire de variables explicatives x transformées ou non

$$y = f(x)$$

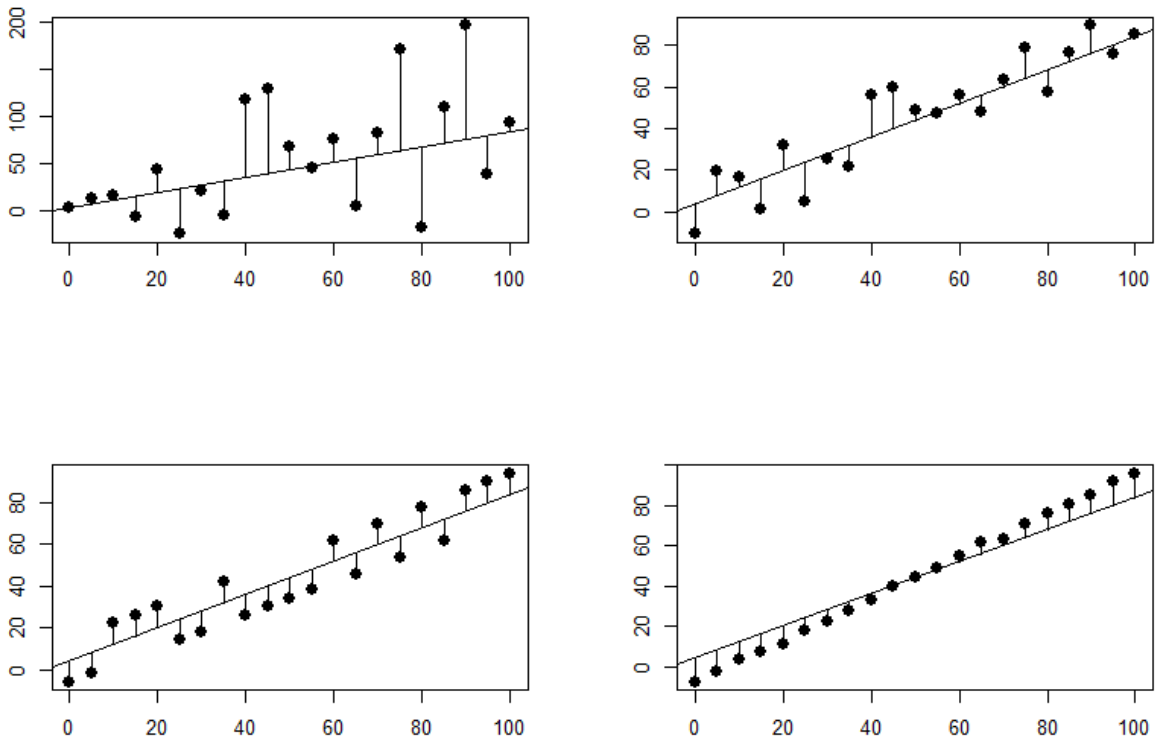


$$y = f(x^2)$$



Qu'est-ce qu'un modèle linéaire ? La résiduelle

$$y = f(x) + e$$

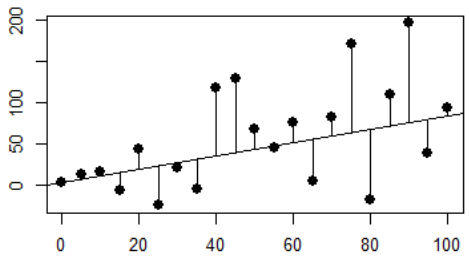


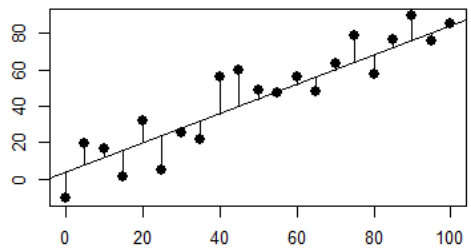
Qu'est-ce qu'un modèle linéaire ? La résiduelle

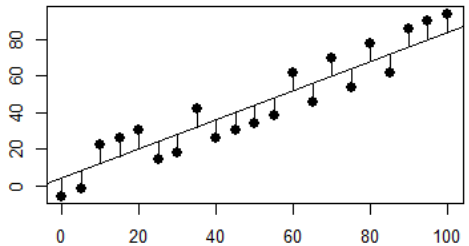
$$y = f(x) + e$$

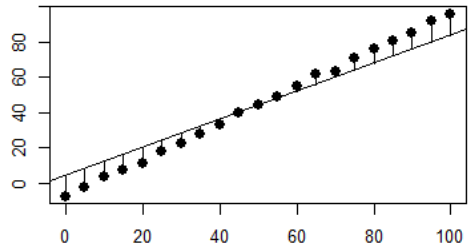
e

$$E(e) = 0$$
$$cov(x, e) = 0$$
$$cov(e_i, e_j) = 0$$







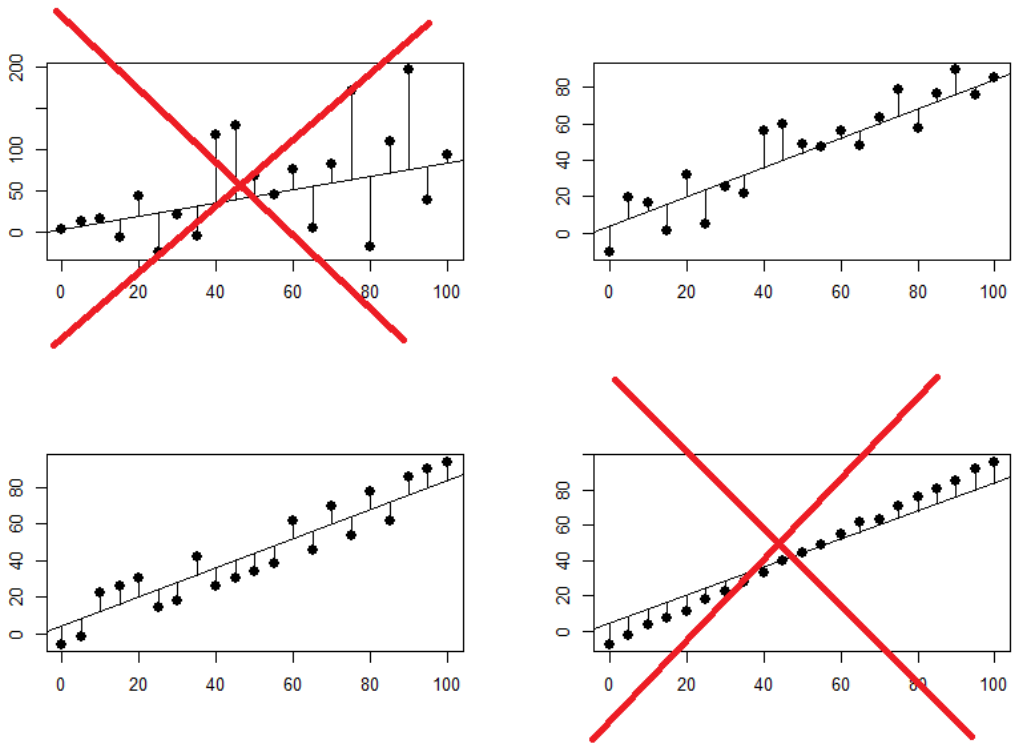


Qu'est-ce qu'un modèle linéaire ? La résiduelle

$$y = f(x) + e$$

e

$$E(e) = 0$$
$$cov(x, e) = 0$$
$$cov(e_i, e_j) = 0$$

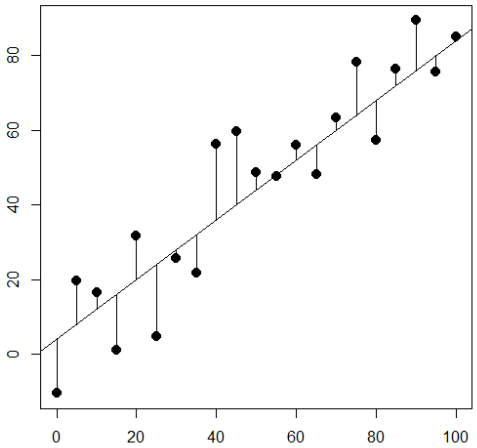


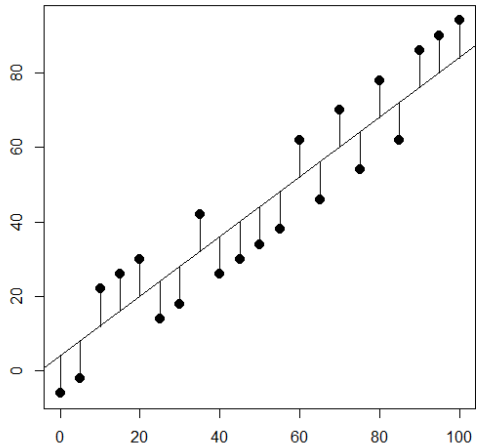
Qu'est-ce qu'un modèle linéaire ? La résiduelle

$$y = f(x) + e$$

e

$$E(e) = 0$$
$$cov(x, e) = 0$$
$$cov(e_i, e_j) = 0$$



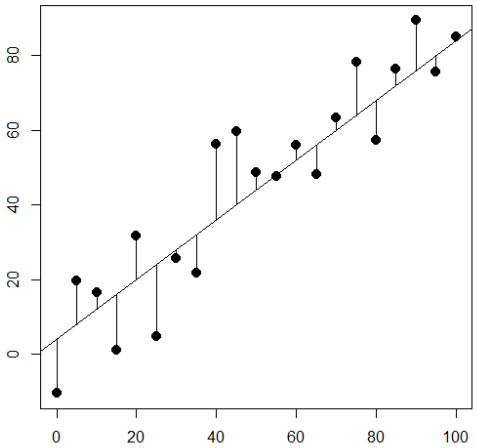


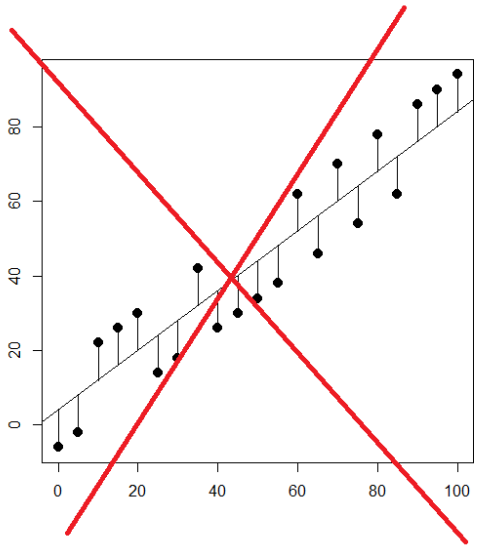
Qu'est-ce qu'un modèle linéaire ? La résiduelle

$$y = f(x) + e$$

e

$$E(e) = 0$$
$$cov(x, e) = 0$$
$$cov(e_i, e_j) = 0$$





Le cadre statistique fréquentiste

- x : régresseur, explicative, prédictive, indépendante, endogène. Les valeurs de cette variable sont fixées
- y : réalisation d'une variable aléatoire gaussienne (normale) $Y \sim N(\mathbb{E}(Y), \sigma^2)$. Y est la variable réponse, expliquée, dépendante, endogène...

- Modèle vrai/déterministe : $\mathbb{E}(Y) = \beta_1 x + \beta_0$
- $\beta_1 et \beta_0$ sont les paramètres
- Variable aléatoire résiduelle : $E \sim N(\mathbb{E}(Y), \sigma^2)$
  - Espérance nulle
  - Variance constante (homoscédasticité)
  - Non corrélés entre eux

Un modèle pourquoi faire : Estimation

- Les valeurs de x sont fixées
- y est la réalisation de la v.a. $Y \sim N(\mathbb{E}(Y), \sigma^2)$, avec $\mathbb{E}(Y) = \beta_1 x + \beta_0$

On veut estimer β_1, β_0 et σ à partir de y. L'**estimateur** est une variable aléatoire qui donne une valeur numérique (**estimation**) à partir de y.

Le cadre statistique fréquentiste

Cadre fréquentiste

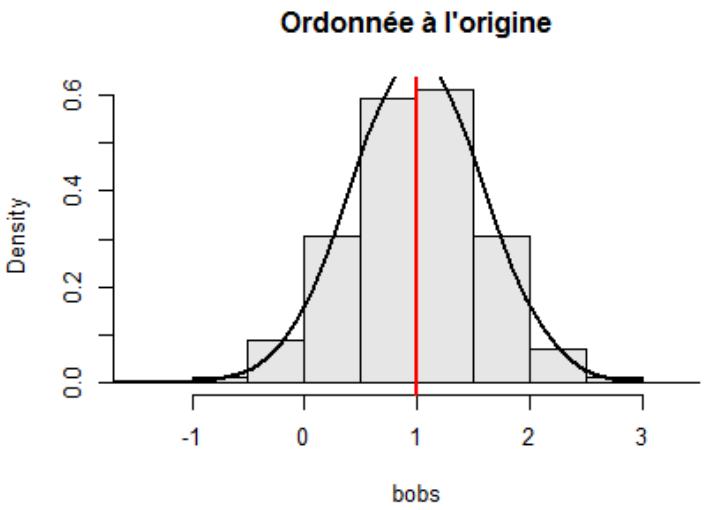
- Les valeurs de x sont fixées
- y est la réalisation de la v.a. $Y \sim N(\mathbb{E}(Y), \sigma^2)$, avec $\mathbb{E}(Y) = \beta_1 x + \beta_0$
- Les estimateurs des β sont des v.a., fonctions de Y
- Les estimations des β, $\hat{\beta}$ sont des fonctions de y

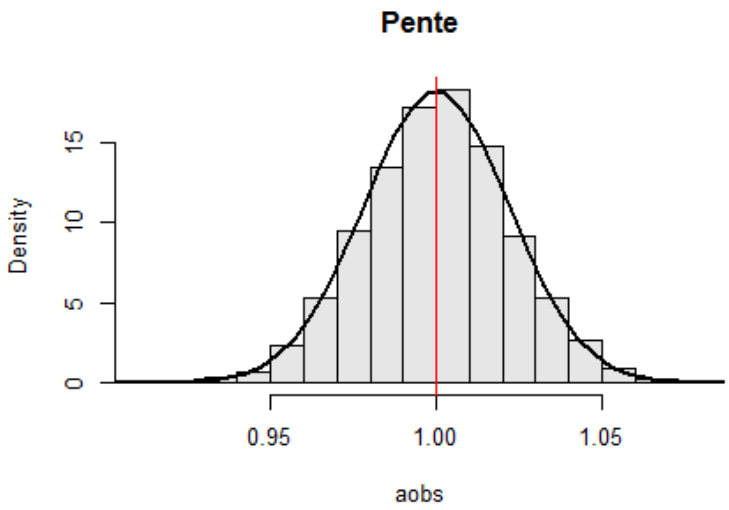
Un exemple

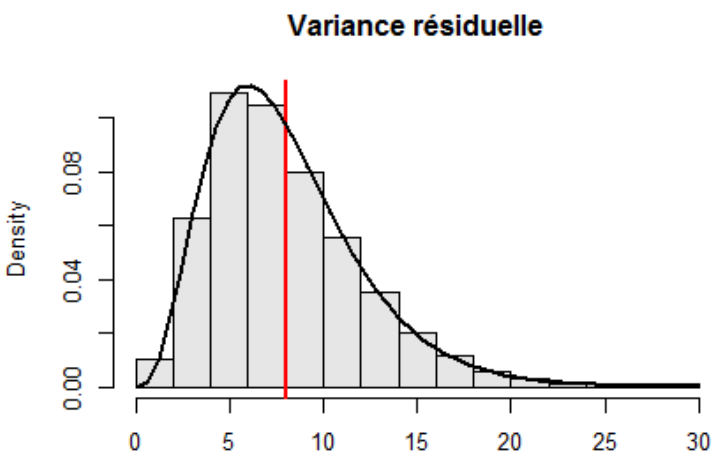
- β_1,β_0 et σ sont fixés à 1
- On simule 5000 échantillons de 10 valeurs à partir de ces paramètres
- Pour chaque échantillon, on calcule les estimations des paramètres
- On représente la distribution empirique des estimations

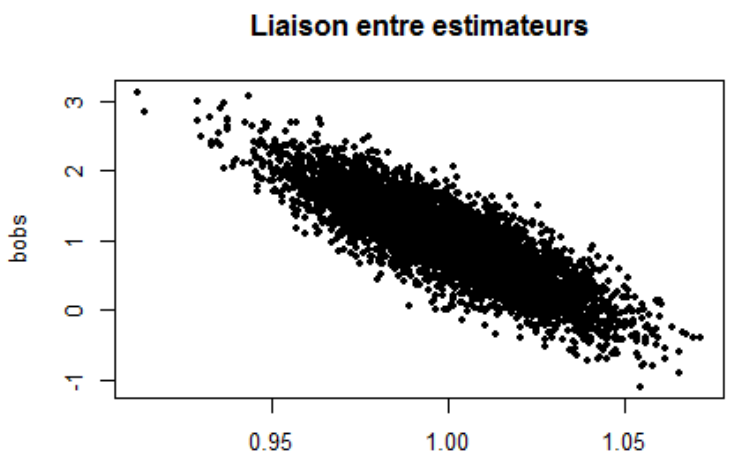
Estimation

Distribution des estimateurs (n=10)



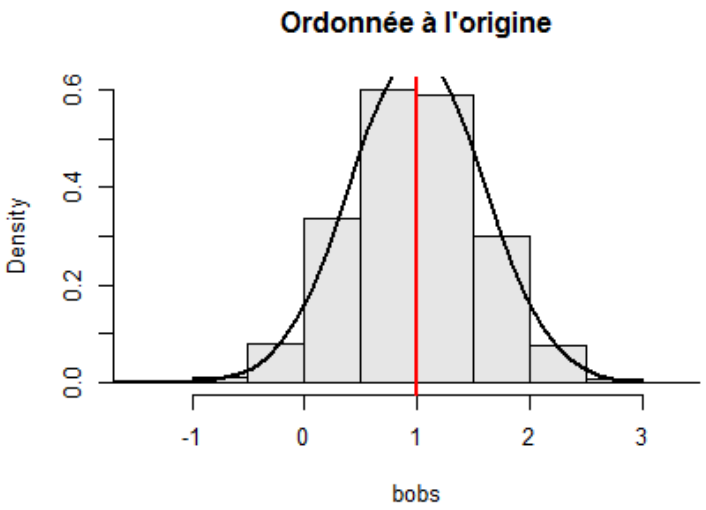


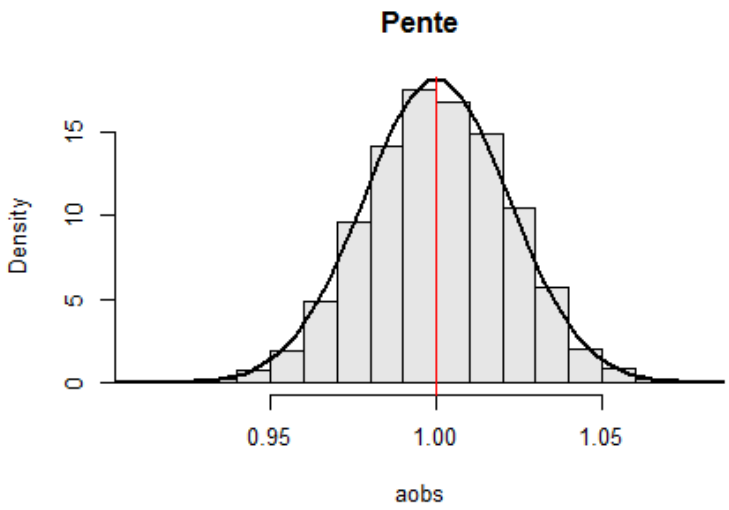


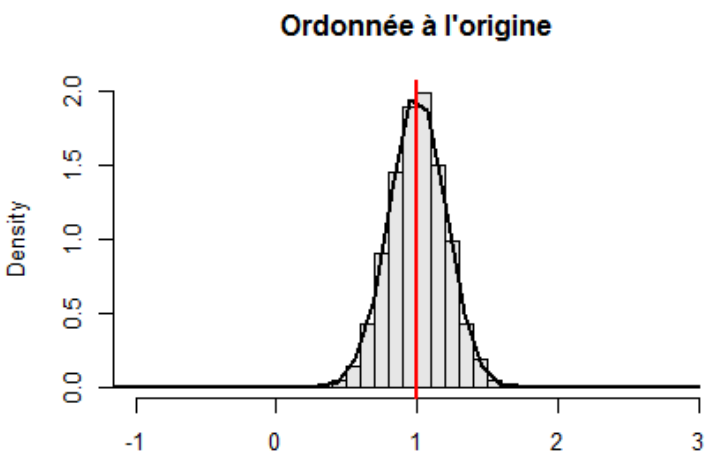


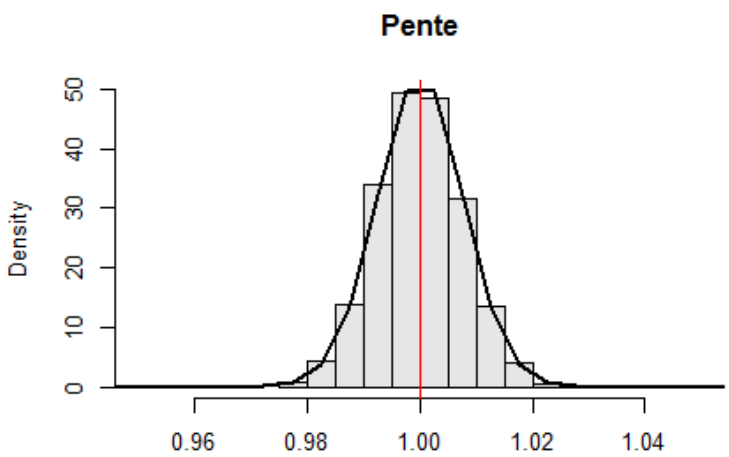
Estimation et taille de l'échantillon

Distribution des estimateurs (n=10 vs n=1000)



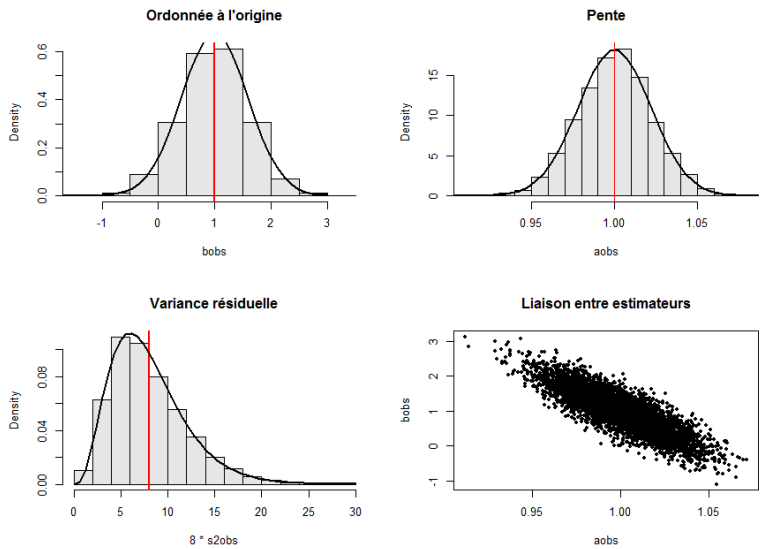






Estimation

- Les estimateurs suivent des distributions, qu'on peut approcher par des distributions classiques, qui dépendent de la taille de l'échantillon
- Covariance entre estimateurs



Propriété des estimateurs

- Biais $\mathbb{E}(\hat{\theta}) - \theta$
- Variance / Erreur quadratique:$\mathbb{E}((\hat{\theta} - \theta)^2)$
- Convergence $\mathbb{P}(\lim_{n \to +\infty} \hat{\theta}_n = \theta) = 1$

Un modèle pourquoi faire ? Prédiction

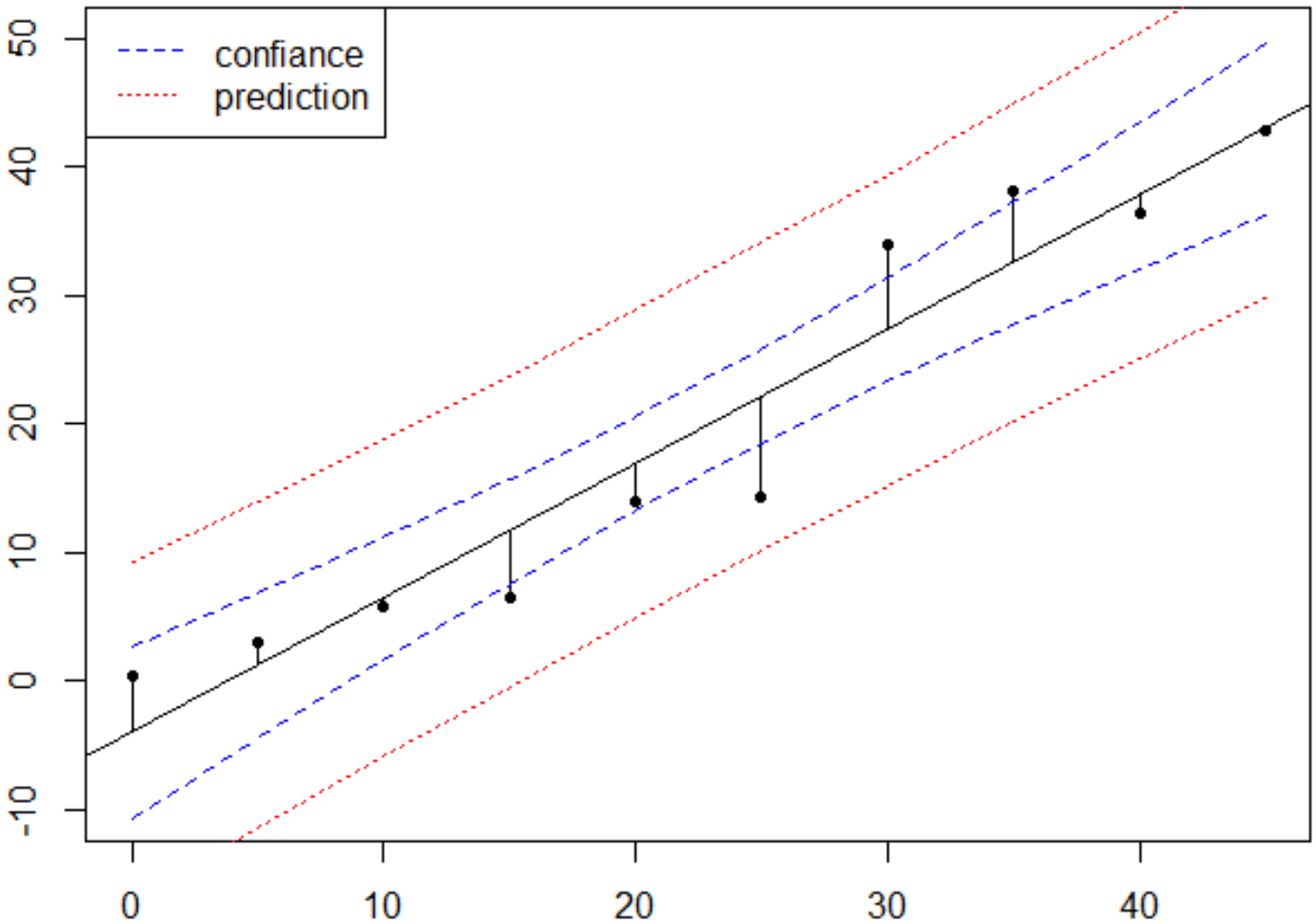
Cadre fréquentiste

- Les valeurs de x sont fixées
- y est la réalisation de la v.a. $Y \sim N(\mathbb{E}(Y), \sigma^2)$, avec $\mathbb{E}(Y) = \beta_1 x + \beta_0$

Prédiction

Est-ce que je peux prédire y pour une (nouvelle) valeur de x

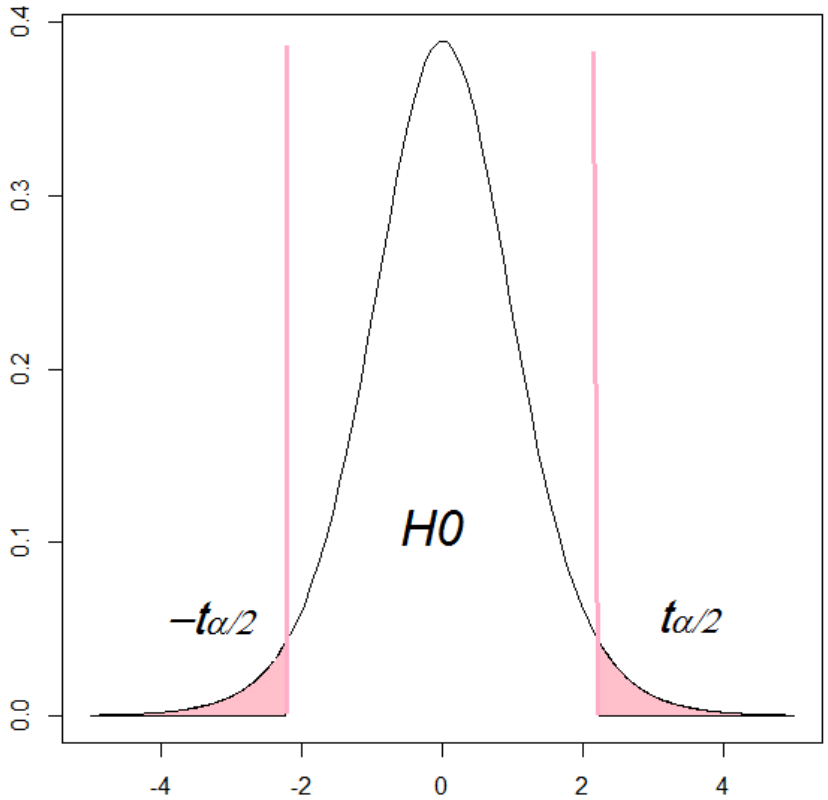
Un modèle pourquoi faire ? Prédiction



Un modèle pourquoi faire ? Tests, hypothèses, inférence

Inférence - Approche de Fisher

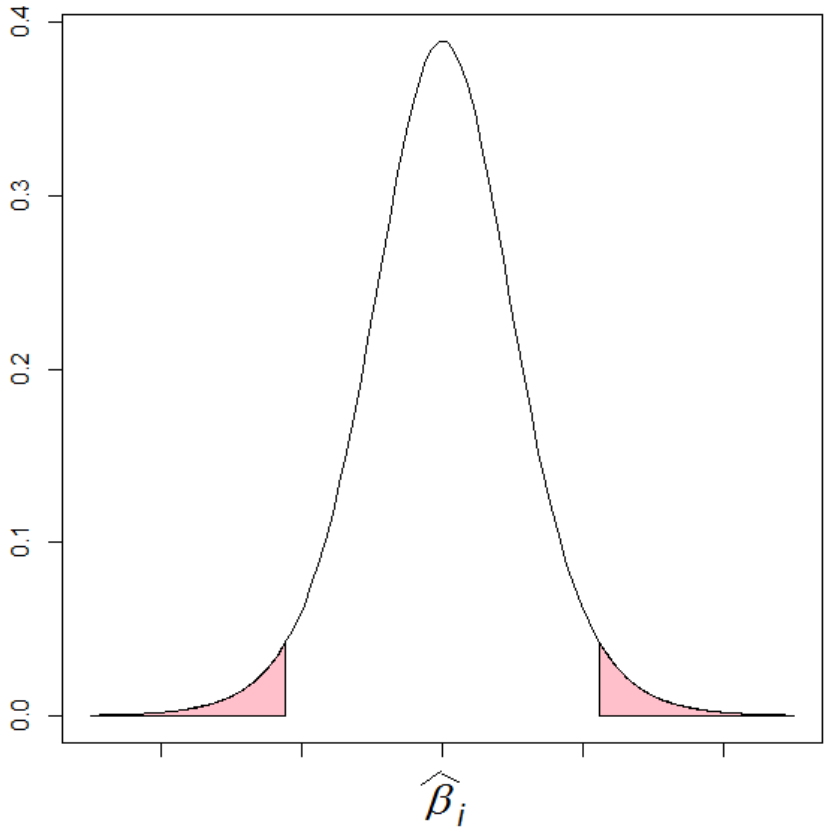
- Une hypothèse à tester
- $\beta = 0$, seuil α



Un modèle pourquoi faire ? Tests, hypothèses, inférence

Inférence - Approche de Fisher

- Intervalle de confiance
- $(1-\alpha)$ % des IC contiennent la vraie valeur de β



Modèle linéaire et type de variables

Données concernant des chats (MASS, cats)

cats {MASS}

Anatomical Data from Domestic Cats

Description

The heart and body weights of samples of male and female cats used for *digitalis* experiments. The cats were all adult, over 2 kg body weight.

Usage

```
cats
```

Format

This data frame contains the following columns:

`Sex`

sex: Factor with evels `"F"` and `"M"`.

`Bwt`

body weight in kg.

`Hwt`

heart weight in g.

Source

R. A. Fisher (1947) The analysis of covariance method for the relation between a part and the whole, *Biometrics* **3**, 65–68.

References

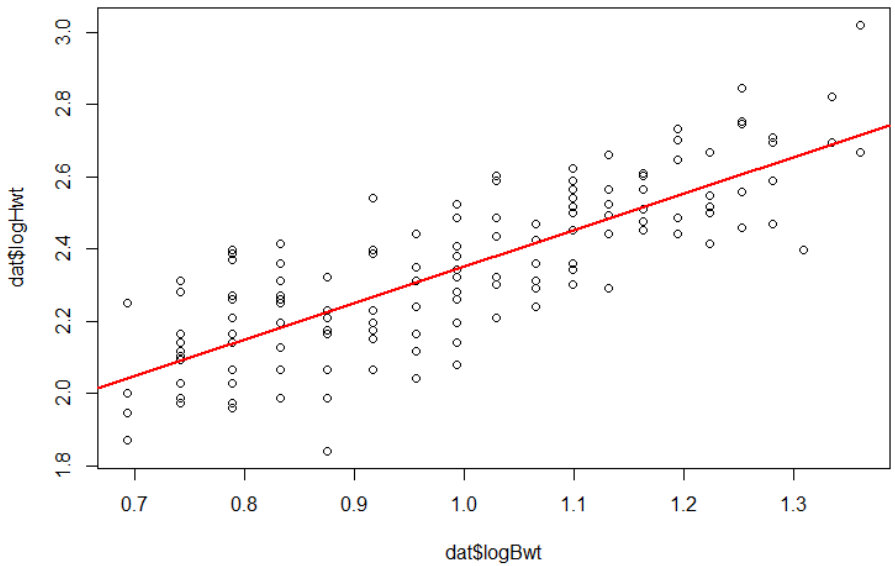
Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.

[Package *MASS* version 7.3-45 Index]

Modèle linéaire et type de variables

Variables quantitatives

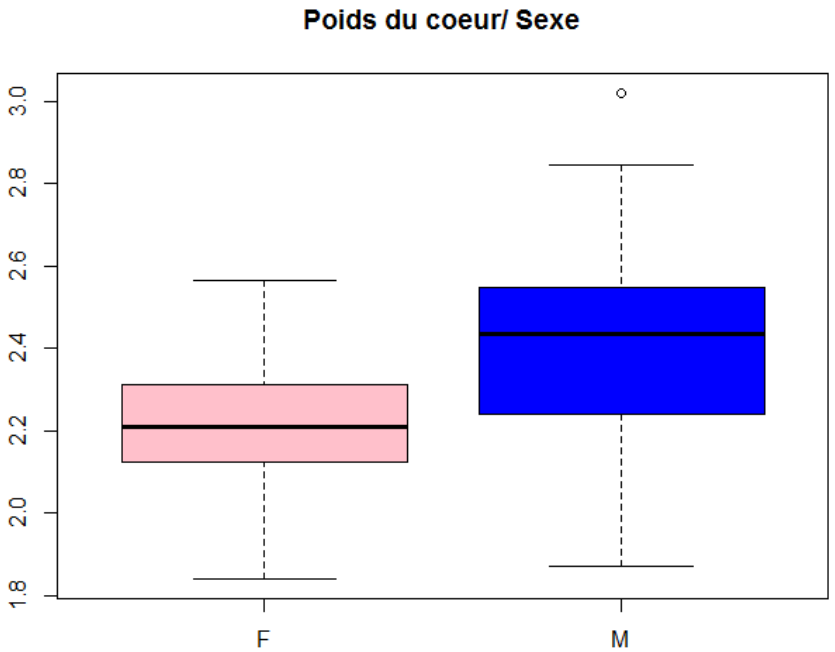
- r=0,80
- Modèle de régression $r^2 = R^2$



Modèle linéaire et type de variables

Variables quantitatives/qualitative

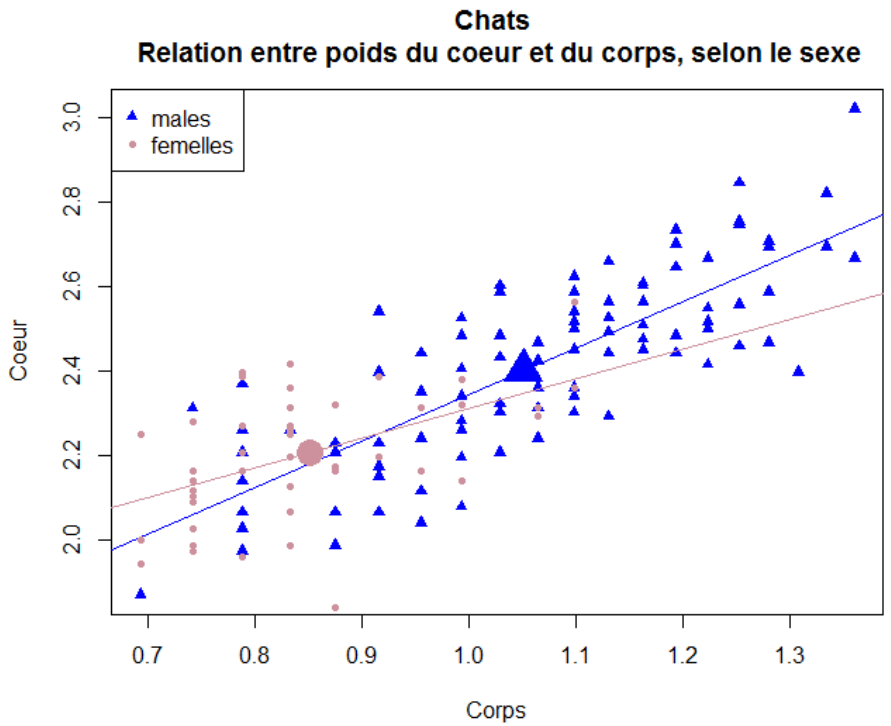
- (Carré du) rapport de corrélation η^2
- Analyse de variance $\eta^2 = R^2$



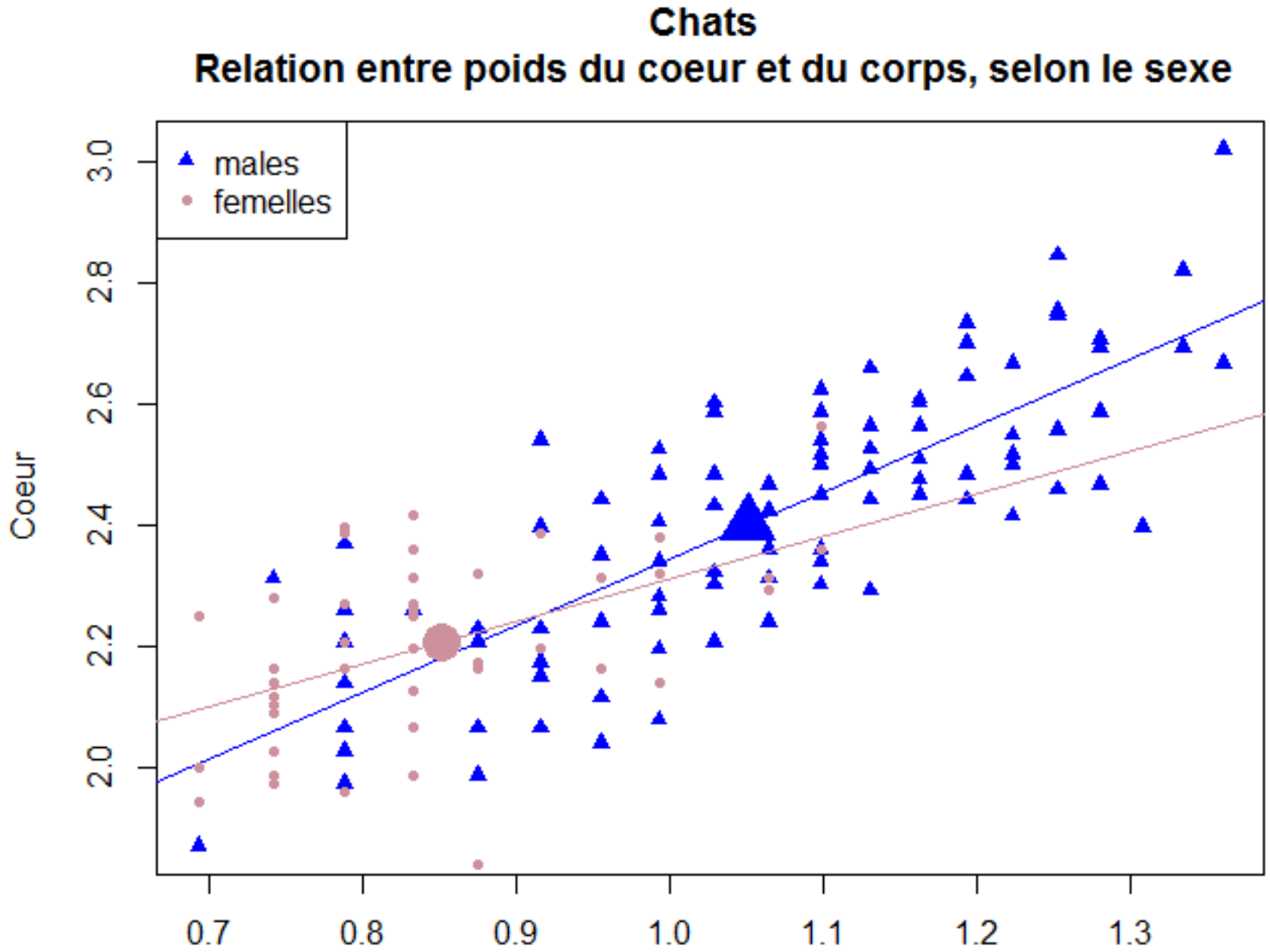
Modèle linéaire et type de variables

Variables quantitatives/quantitative + qualitative

- Analyse de covariance



Analyse de covariance



En résumé

Modélisation statistique

- une variable réponse Y est modélisée par des combinaisons linéaires de variables explicatives x (ou leur transformées)
- Une partie fixe : $\mathbb{E}(Y) = \sum \beta_i x_i$
- un résidu aléatoire : $E \sim N(0, \sigma)$

y réalisation d'une variable aléatoire Y

- Estimation des β
- L'estimateur de β est une v.a., dont on a une réalisation, l'estimation.
- Tests d'hypothèse, IC
- Prédiction

Nature des variables explicatives

- Quantitative : régression
- Qualitatif : analyse de variance
- Quantitatif+Qualitatif : analyse de covariance