

Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #1 Régression (introduction générale)

Régression Linéaire

Un modèle de régression implique:

- ▶ une variable y ($y \in \mathbb{R}$, ou un sous-espace) appelée: variable à expliquer, endogène, dépendante (donnée)
- ▶ des variables x_1, \dots, x_p ($p \geq 1$) appelées: covariables, prédicteurs, variables exogènes, facteurs (données)
- ▶ un vecteur de paramètres $\beta \in \mathbb{R}^q$ ($q \geq 0$) (à estimer)

Formellement

- ▶ un modèle de régression consiste en $y \approx f(x_1, \dots, x_p, \beta)$
- ▶ on suppose que les $(y_i, x_{1,i}, \dots, x_{p,i})$ sont des réalisations du vecteur aléatoire (Y, X_1, \dots, X_p)
- ▶ et on suppose que $\mathbb{E}(Y \mid x_1, \dots, x_p) = f(x_1, \dots, x_p, \beta)$.

Régression Linéaire (et autres)

- ▶ *Régression linéaire (multiple)*: $Y \in \mathbb{R}$,
 $f(x_1, \dots, x_p; \beta) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$,
- ▶ *Régression polynomiale*: $Y \in \mathbb{R}$,
 $f(x_1; \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \dots + \beta_s x_1^s$.
- ▶ *Régression logistique*: $Y \in \{0, 1\}$,
 $\text{logit} \mathbb{E}(Y \mid x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.
- ▶ *Rég. Poisson*: $Y \in \mathbb{N}$,
 $\log \mathbb{E}(Y \mid x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.
- ▶ *Modèles mixtes*: combine des effets fixes et aléatoires permettant de prendre en compte des données longitudinales.
- ▶ *Régression. non paramétrique simple*: $Y = m(x_1) + \varepsilon$,
 $m : \mathbb{R} \rightarrow \mathbb{R}$.
- ▶ *Régression non paramétrique multiple*: $Y = m(x_1, \dots, x_p) + \varepsilon$,
 $m : \mathbb{R}^p \rightarrow \mathbb{R}$.
- ▶ *Modèles additifs généralisés*: $Y = m_1(x_1) + \dots + m_p(x_p) + \varepsilon$,
 $m_j : \mathbb{R} \rightarrow \mathbb{R}$.

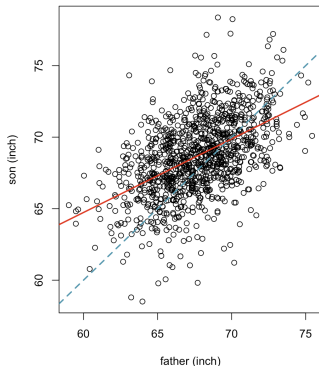
▶ ...

Exemple (régression linéaire simple)

```
1 > library(UsingR)
2 > head(father.son)
3   fheight  sheight
4 1 65.04851 59.77827
5 2 63.25094 63.21404
6 3 64.95532 63.34242
7 4 65.75250 62.79238
8 5 61.13723 64.28113
9 6 63.02254 64.24221
```

si y_i et x_i les tailles des fils et père de la famille i ,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$



Exemple (Régression logistique)

```
1 > loc = "http://freakonometrics.free.fr/titanic.RData"
2 > download.file (loc , "titanic.RData")
3 > load ("titanic.RData")
4 > str(titanic)
5 'data.frame': 2207 obs. of  9 variables:
6 $ gender   : Factor w/ 2 levels "female","male": 2 ...
7 $ age      : num  42 13 16 39 16 25 30 28 27 20 ...
8 $ class    : Factor w/ 7 levels "1st","2nd","3rd",...
9 $ embarked: Factor w/ 4 levels "Belfast","Cherbourg"
10 $ country  : Factor w/ 48 levels "Argentina", ...
11 $ fare     : num  7.11 20.05 20.05 20.05 7.13 ...
12 $ sibsp    : num  0 0 1 1 0 0 1 1 0 0 ...
13 $ parch    : num  0 2 1 1 0 0 0 0 0 0 ...
14 $ survived: Factor w/ 2 levels "no","yes": 1 1 ...
```

Exemple (Régression non paramétrique)

Jeu de données standard `car::prestige`; 102 observations; 6 covariables décrivant le prestige d'emplois occupés; recensement Canadien en 1971.

- ▶ `education`: éducation moyenne;
- ▶ `income`: revenu moyen en \$;
- ▶ `women`: % de femmes;
- ▶ `prestige`: Pineo-Porter prestige score de l'emploi (provenant d'une étude sociale conduite dans les années 60);
- ▶ `census`: code de l'emploi;
- ▶ `type`: type de l'emploi (`bc`, Blue Collar; `prof`, Professional, Managerial, and Technical; `wc`, White Collar).

Exemple (Régression non paramétrique)

```
1 > library(car)
2 > head

```

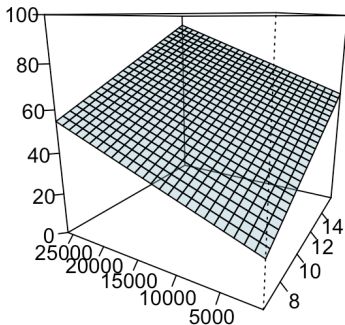
prestige)
3
4 education income women prest cens type
5 gov.administrators 13.11 12351 11.16 68.8 1113 prof
6 general.managers 12.26 25879 4.02 69.1 1130 prof
7 accountants 12.77 9271 15.70 63.4 1171 prof
8 purchasing.officers 11.42 8865 9.11 56.8 1175 prof
9 chemists 14.62 8403 11.68 73.5 2111 prof
10 physicists 15.64 11030 5.13 77.6 2113 prof
```


```

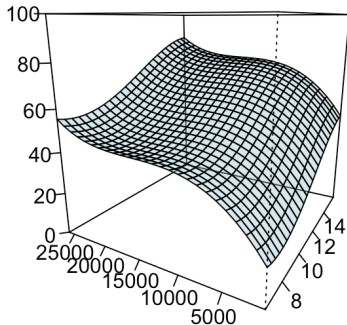
- ▶ Intéressons-nous à la relation entre income et prestige.
- ▶ Si aucune relation claire, une approche **non paramétrique** peut être pertinente.
- ▶ **Modèle:** $\text{prestige}_i = m(\text{income}_i) + \varepsilon_i$, $m(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ est le paramètre fonctionnel à estimer.

Exemple (Régression non paramétrique)

$$\beta_0 + \beta_1 \text{income}_i + \beta_2 \text{education}_i + \varepsilon_i$$
$$(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$$



$$m(\text{income}_i, \text{education}_i) + \varepsilon_i$$
$$m : \mathbb{R}^2 \rightarrow \mathbb{R}$$



One can consider some additive model, $m_1, m_2 : \mathbb{R} \rightarrow \mathbb{R}$,
 $\text{prestige}_i = \beta_0 + m_1(\text{income}_i) + m_2(\text{education}_i) + \varepsilon_i$,

Exemple (Données)

Prices of houses (per m^2) in Warsaw (Poland)

```
1 > library(DALEX)
2 > data(apartments)
3 > str(apartments)
4 'data.frame': 1000 obs. of  6 variables:
5  $ m2.price           : num  5897 1818 3643 3517 ...
6  $ construction.year : num  1953 1992 1937 1995 ...
7  $ surface            : num  25 143 56 93 144 61 ...
8  $ floor              : int   3 9 1 7 6 6 8 8 6 9 ...
9  $ no.rooms           : num   1 5 2 3 5 2 5 4 6 4 ...
10 $ district           : Factor w/ 10 levels "Bemowo" ...
```

Exemple (Données)

Rent of houses in Munich (Germany)

```
1 > loc = "http://freakonometrics.free.fr/rent98_00.txt"
2 > munich = read.table(loc,header=TRUE)
3 > str(munich)
4 'data.frame': 4571 obs. of 13 variables:
5 $ rent_euro : num 121 436 355 282 805 ...
6 $ rentsqm_euro: num 3.45 4.19 12.23 7.23 8.3 ...
7 $ area : int 35 104 29 39 97 62 31 61 72 ...
8 $ yearc : num 1939 1939 1971 1972 1985 ...
9 $ glocation : int 0 0 1 1 0 0 0 1 0 0 ...
10 $ tlocation : int 0 0 0 0 0 0 0 0 0 0 ...
11 $ nkitchen : int 0 0 0 0 0 0 0 0 0 0 ...
12 $ pkitchen : int 0 0 0 0 0 0 0 0 0 0 ...
13 $ eboden : int 1 0 1 1 0 0 1 0 0 0 ...
14 $ year01 : int 0 0 0 0 0 0 0 0 0 0 ...
15 $ yearc2 : num ...
16 $ yearc3 : num ...
17 $ invarea : num 0.02857 0.00962 0.03448 ...
```

Exemple (Données)

```
1 > Davis = read.table(  
2 "http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-  
   Regression-2E/datasets/Davis.txt")  
3 > Davis[12,c(2,3)] = Davis[12,c(3,2)]  
4 > str(Davis)  
5 'data.frame': 200 obs. of 5 variables:  
6 $ sex : Factor w/ 2 levels "F","M": 2 1 ...  
7 $ weight : int 77 58 53 68 59 76 76 69 71 ...  
8 $ height : int 182 161 161 177 157 170 167...  
9 $ reportedWeight: int 77 51 54 70 59 76 77 73 71 ...  
10 $ reportedHeight: int 180 159 158 175 155 165 165...
```