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 $x_1, \ldots, x_p$ ($p \geq 1$) appelées: covariables, prédicteurs, variables exogènes, facteurs (données)
- un vecteur de paramètres $\boldsymbol{\beta} \in \mathbb{R}^q$ ($q \geq 0$) (à estimer)

Formellement

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

Régression Linéaire (et autres)

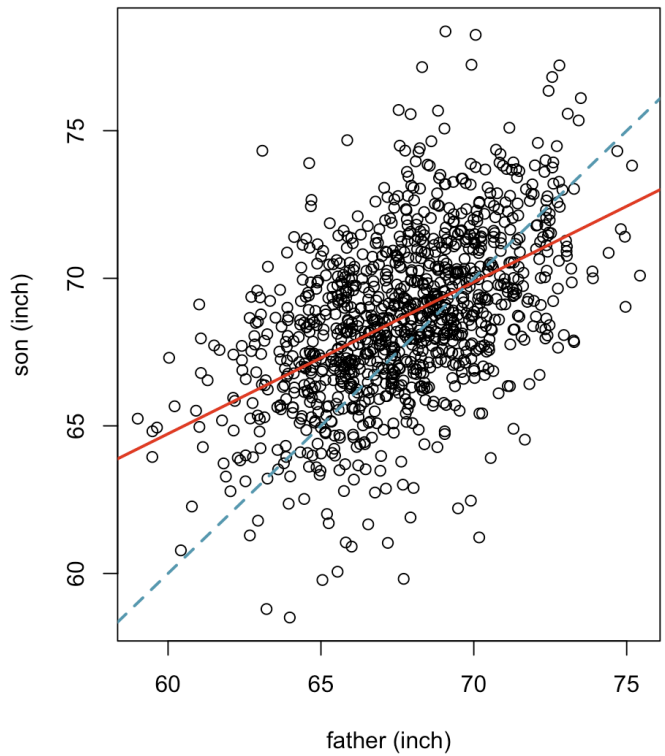
- *Régression linéaire (multiple)*: $Y \in \mathbb{R}$,
  $f(x_1, \ldots, x_p; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$,
- *Régression polynomiale*: $Y \in \mathbb{R}$,
  $f(x_1; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \beta_s x_1^s$.
- *Régression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

Exemple (régression linéaire simple)

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

```
> loc = "http://freakonometrics.free.fr/titanic.RData"
> download.file (loc , "titanic.RData")
> load ("titanic.RData")
> str(titanic)
'data.frame': 2207 obs. of  9 variables:
 $ gender  : Factor w/ 2 levels "female","male": 2 ...
 $ age     : num  42 13 16 39 16 25 30 28 27 20 ...
 $ class   : Factor w/ 7 levels "1st","2nd","3rd",..
 $ embarked: Factor w/ 4 levels "Belfast","Cherbourg"
 $ country : Factor w/ 48 levels "Argentina", ...
 $ fare    : num  7.11 20.05 20.05 20.05 7.13 ...
 $ sibsp   : num  0 0 1 1 0 0 1 1 0 0 ...
 $ parch   : num  0 2 1 1 0 0 0 0 0 0 ...
 $ 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 étudie 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).

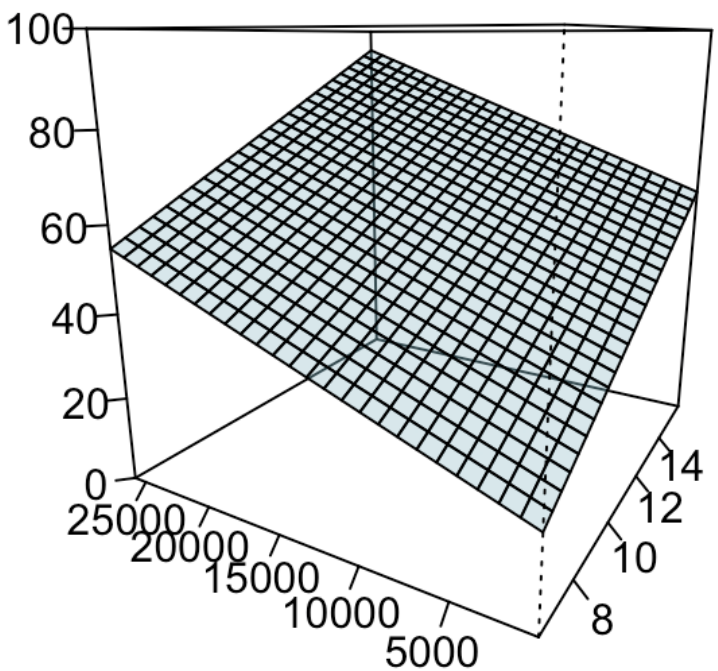
Exemple (Régression non paramétrique)

```
> library(car)
> head(prestige)
                    education income women prest cens type
gov.administrators  13.11     12351  11.16  68.8 1113 prof
general.managers    12.26     25879   4.02  69.1 1130 prof
accountants         12.77      9271  15.70  63.4 1171 prof
purchasing.officers 11.42      8865   9.11  56.8 1175 prof
chemists            14.62      8403  11.68  73.5 2111 prof
physicists          15.64     11030   5.13  77.6 2113 prof
```

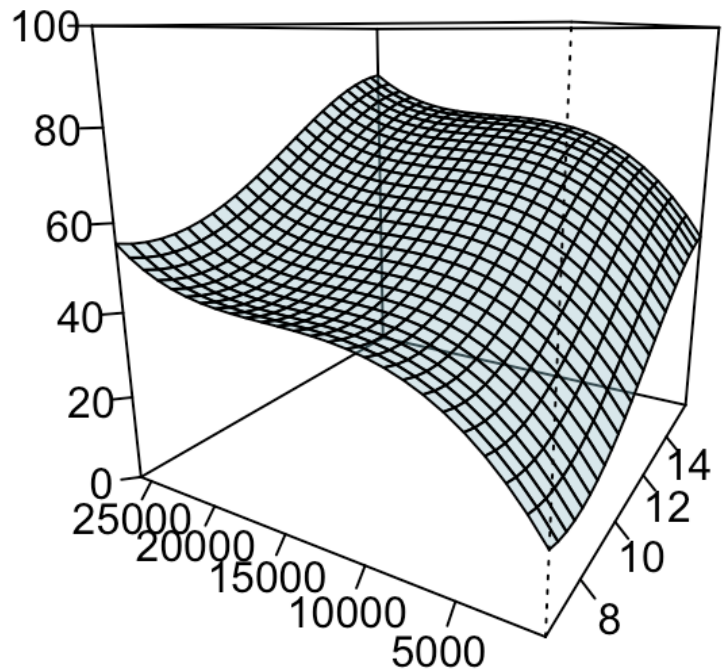
- Intéressons-nous à la relation entre income et prestige.
- Si aucune relation claire, une approche **non paramétrique** peut être pertinente.
- **Modèle**: $\text{prestige}_i = m(\text{income}_i) + \varepsilon_i$, $m(\cdot) : \mathbb{R} \to \mathbb{R}$ est le paramètre fonctionnel à estimer.

Exemple (Régression non paramétrique)

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



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



One can consider some additive model, $m_1, m_2 : \mathbb{R} \to \mathbb{R}$,

$$\text{prestige}_i = \beta_0 + m_1(\text{income}_i) + m_2(\text{education}_i) + \varepsilon_i,$$

Exemple (Données)

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

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

Exemple (Données)

Rent of houses in Munich (Germany)

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

Exemple (Données)

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