Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2Q20

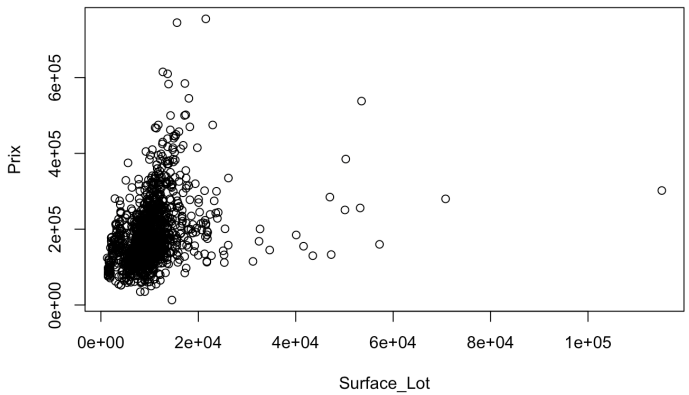
OLS #23 (example)

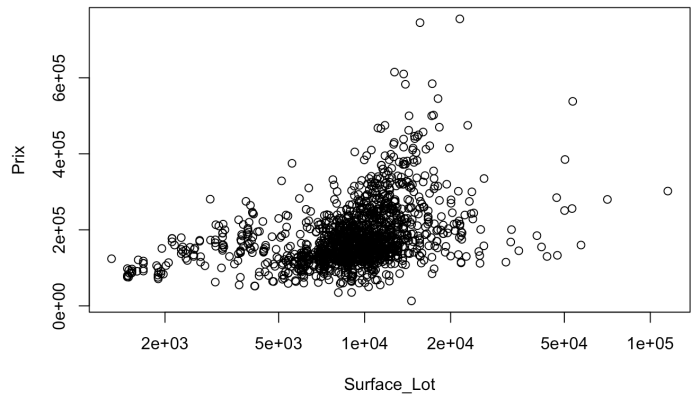
Moindres carrés

```
> loc_fichier = "http://freakonometrics.free.fr/prix_maison.RData"
> download.file(loc_fichier, "base2.RData")
> load("base2.RData")
> dim(database)
[1] 1427   41
> str(database)
'data.frame': 1427 obs. of  41 variables:
 $ Zone               : Factor w/ 7 levels "A (agr)","C (all)",..: 6 6 6 6 6 ...
 $ Surface_Lot        : int  8396 11631 10456 14694 10400 9760 2998 6000 7400 ...
 $ Rue                : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
 $ Forme              : Factor w/ 4 levels "IR1","IR2","IR3",..: 1 1 1 1 4 4 ...
 $ Utilities          : Factor w/ 3 levels "AllPub","NoSeWa",..: 1 1 1 1 1 1 ...
 $ Configuration      : Factor w/ 5 levels "Corner","CulDSac",..: 5 1 5 5 5 5 ...
 $ Proxim_1           : Factor w/ 9 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 ...
 $ Proxim_2           : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 ...
 $ Logement           : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 5 ...
 $ Style              : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 6 3 3 6 6 ...
 $ Int_Qualite        : int  7 8 6 8 6 6 6 6 7 4 ...
 $ Int_Condition      : int  5 5 6 9 5 8 5 7 6 7 ...
 $ Construction_Annee: int  2003 2004 1967 1977 1972 1964 2000 1940 1962 ...
 $ Renovation_Annee   : int  2003 2005 1967 2008 1972 1993 2000 1989 1962 ...
 $ Toit               : Factor w/ 6 levels "Flat","Gable",..: 2 2 4 2 2 4 2 ...
 $ Exterieur          : Factor w/ 16 levels "AsbShng","AsphShn",..
 $ Maconnerie         : Factor w/ 6 levels "","BrkCmn","BrkFace",..
 $ Ext_Qualite        : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 4 1 4 4 3 ...
 $ Ext_Condition      : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 1 5 5 5 ...
 $ Foundation         : Factor w/ 6 levels "BrkTil","CBlock",..: 3 3 2 2 2 ...
 $ Chauffage          : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 ...
 $ Chauff_Qualite     : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 3 1 3 3 1 ...
 ...
 $ Garage_Int_Surface: int  0 0 0 0 0 0 0 0 0 98 ...
 $ Prix               : int  213000 258000 218500 318750 165150 ...
```

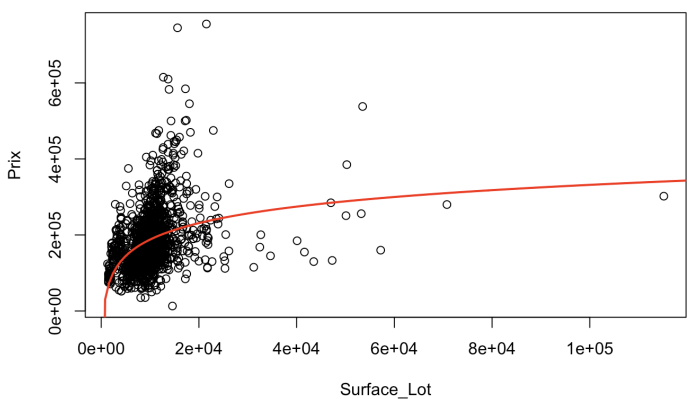
Moindres carrés

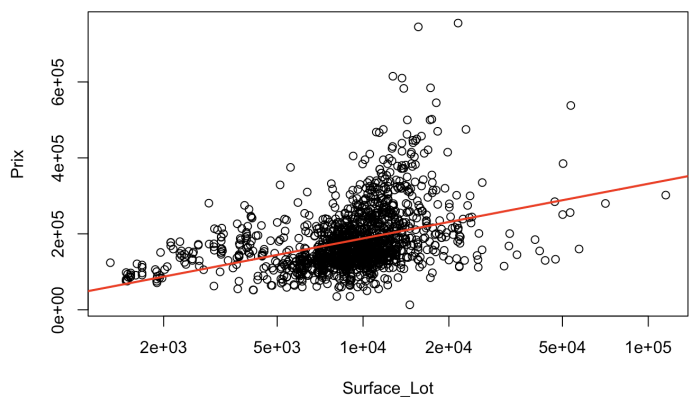
```
1 with(database,plot(Surface_Lot,Prix))
2 with(database,plot(Surface_Lot,Prix,log="x"))
```





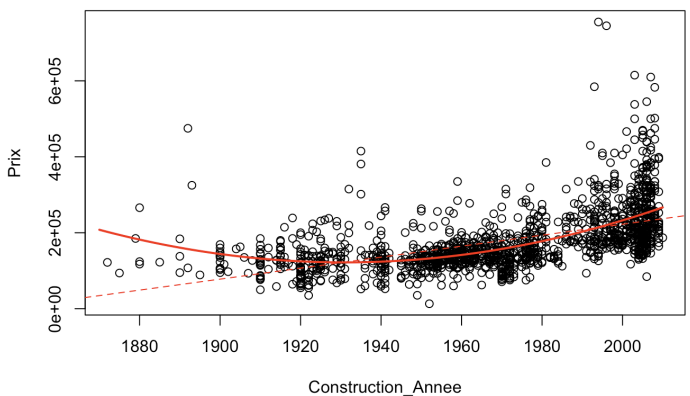

```
1 reg1 = lm(Prix~log(Surface_Lot),data=database)
2 u = seq(0.001,2e+5,length=251)
3 v = predict(reg1,newdata = data.frame(Surface_Lot=u))
4 lines(u,v,lwd=2,col="red")
```

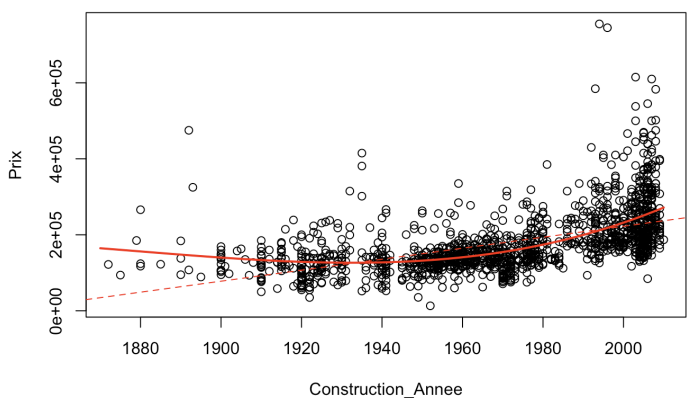




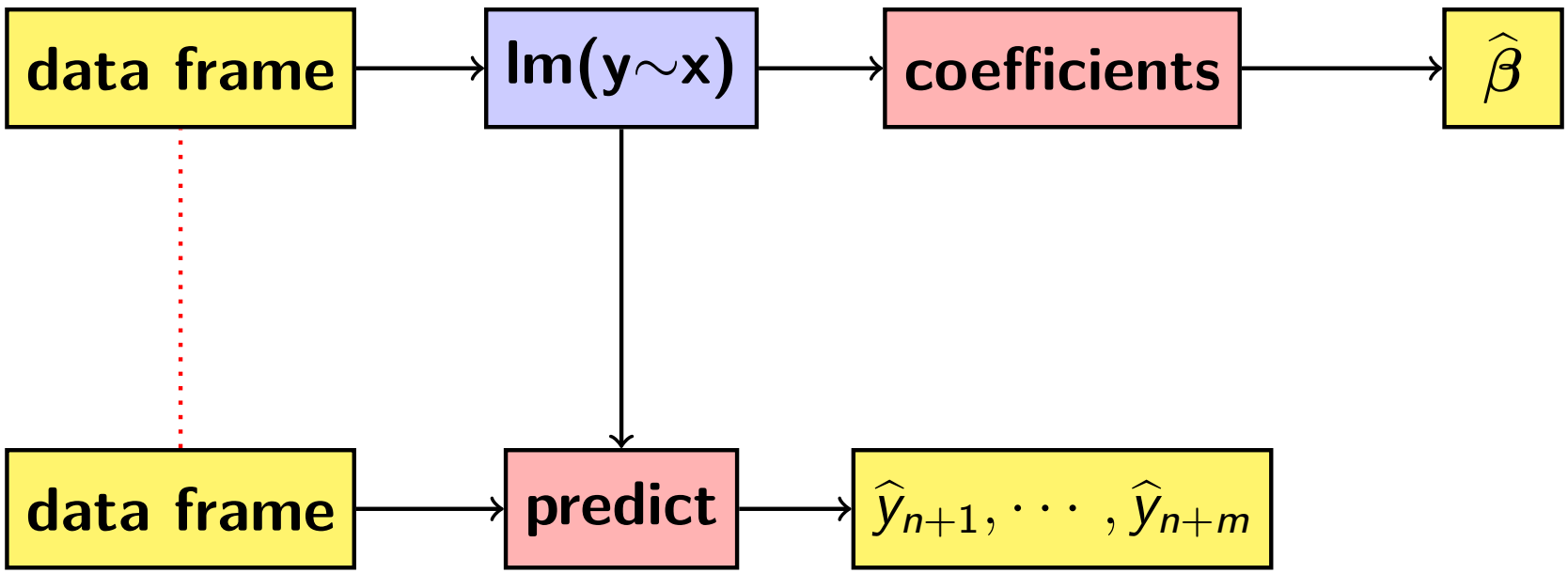
Moindres carrés

```
1 with(database,plot(Construction_Annee,Prix))
2 library(splines)
3 regs = lm(Prix~bs(Construction_Annee),data=database)
4 u=1870:2010
5 v = predict(regs,newdata=data.frame(Construction_Annee = u))
6 lines(u,v,col="red",lwd=2)
7 abline(lm(Prix~(Construction_Annee),data=database),lty=2,col="red")
8 regp = lm(Prix~poly(Construction_Annee,2),data=database)
```





Moindres carrés

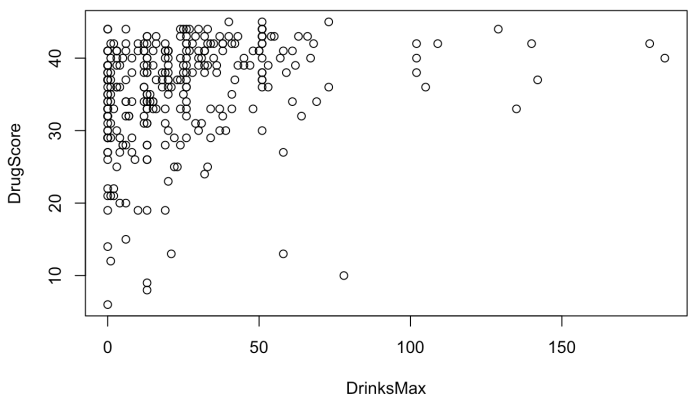


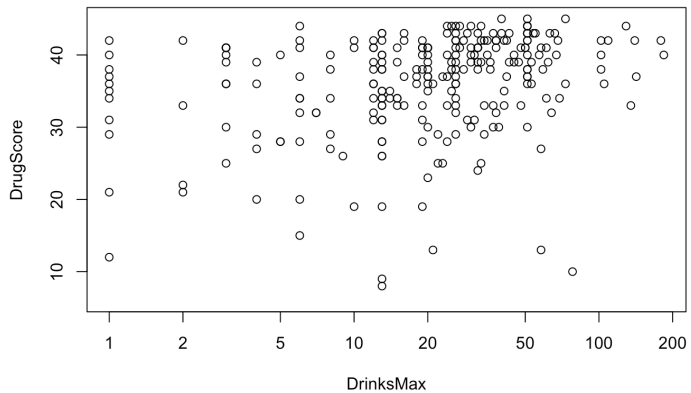
Moindres carrés

```
> loc_fichier = "http://freakonometrics.free.fr/drug_score.RData"
> download.file(loc_fichier, "base3.RData")
> load("base3.RData")
> dim(database)
[1] 300  20
> str(database)
'data.frame': 300 obs. of  20 variables:
 $ Age              : int  32 41 38 28 39 47 34 31 53 27 ...
 $ Depression       : int  39 28 34 36 54 6 38 28 57 52 ...
 $ Hospitalizations: int  0 1 1 1 6 1 2 1 4 0 ...
 $ Link             : int  14 104 348 414 64 365 365 123 365 49 ...
 $ RiskDrug         : int  0 0 14 0 0 0 8 1 0 10 ...
 $ Gender           : Factor w/ 2 levels "female","male": 2 2 1 2 2 1 1 2  ...
 $ Suicide          : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
 $ Homeless         : Factor w/ 2 levels "homeless","housed": 1 2 1 1 1 2 2  ...
 $ DrinksAverage    : int  6 22 0 6 68 4 0 5 38 9 ...
 $ DrinksMax        : int  13 22 0 12 68 4 0 25 51 24 ...
 $ PostCare         : int  1 1 0 0 1 0 0 1 0 1 ...
 $ MentalScore      : num  19.3 39.5 43.4 24.1 13.4 ...
 $ PhysicalScore    : num  59.9 28.9 21.9 52.6 42.1 ...
 $ SocialScore      : int  6 7 9 4 7 5 10 10 5 13 ...
 $ Ethnicity        : Factor w/ 4 levels "black","hispanic",..: 1 1 1 1 4 1  ...
 $ BSAS             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ SexRisk          : int  5 3 8 7 3 5 4 8 2 3 ...
 $ SubstenceAbuse   : Factor w/ 3 levels "alcohol","cocaine",..: 2 1 3 2 2 2 ...
 $ Treated          : Factor w/ 2 levels "no","yes": 2 2 1 2 1 2 1 2 1 2 ...
 $ DrugScore        : int  33 25 32 39 42 29 33 39 45 37 ...
```

Moindres carrés

```
with(database,plot(DrinksMax,DrugScore))
with(database,plot(DrinksMax,DrugScore,log="x"))
```

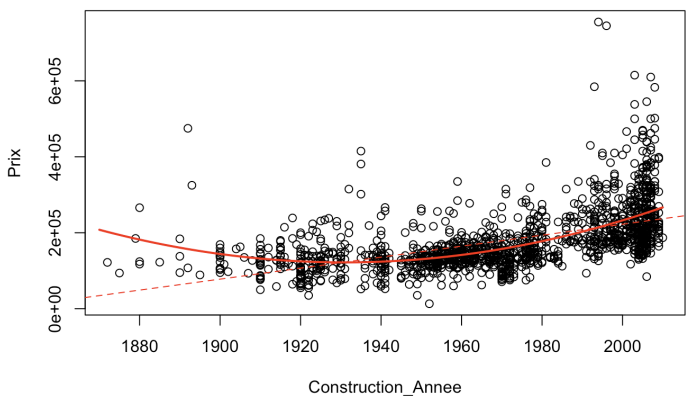


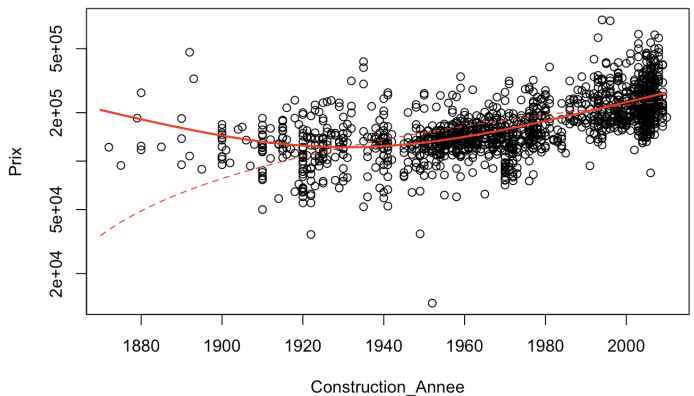



```
modele = lm(DrugScore ~ Homeless+DrinksMax+Gender+MentalScore, data=database)
f = function(x) ifelse(x==0,0,log(x))
modele = lm(DrugScore ~ Homeless+I(DrinksMax==0)+f(DrinksMax)+Gender+MentalScore
     , data=database)
```

Moindres carrés

```
1 with(database,plot(Construction_Annee,Prix))
2 with(database,plot(Construction_Annee,Prix,log="y"))
```





Linear model, approximate $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$ by $\boldsymbol{x}^\top\beta$

Problem: $\mathbb{E}[\log Y|\boldsymbol{X}=\boldsymbol{x}]=\boldsymbol{x}^\top\beta \not\to \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]=\exp[\boldsymbol{x}^\top\beta]$
cf Jensen inequality, $\mathbb{E}(h(Y))\neq h(\mathbb{E}(Y))$
but, in the case of a logarithmic transformation,
$\mathbb{E}[\log Y|\boldsymbol{X}=\boldsymbol{x}]=\boldsymbol{x}^\top\beta \ \to \ \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]=\exp\left[\boldsymbol{x}^\top\beta+\frac{\sigma^2}{2}\right]$