Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2Q20

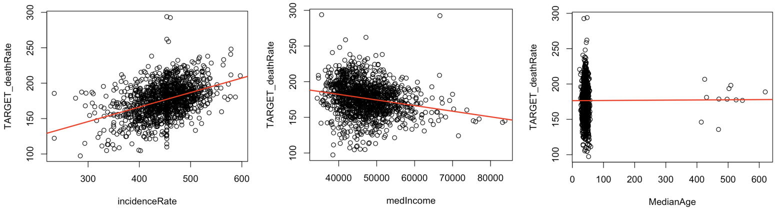
OLS #22 (example)

Moindres carrés

```
> loc_fichier = "http://freakonometrics.free.fr/deathRate.RData"
> download.file(loc_fichier, "deathRate.RData")
> load("deathRate.RData")
> str(database)
'data.frame': 1282 obs. of  32 variables:
 $ avgAnnCount          : num  173 102 427 57 428 ...
 $ avgDeathsPerYear     : int  70 50 202 26 152 71 1380 36 26 901 ...
 $ TARGET_deathRate     : num  161 175 195 144 176 ...
 $ incidenceRate        : num  412 350 430 350 505 ...
 $ medIncome            : int  48127 49348 44243 49955 52313 40189 60397 ...
 $ popEst2015           : int  43269 21026 75882 10321 61023 20848 843954 ...
 $ povertyPercent       : num  18.6 14.6 17.1 12.5 15.6 17.8 13.1 12.7 12.6 ...
 $ studyPerCap          : num  23.1 47.6 342.6 0 180.3 ...
 $ binnedInc            : Factor w/ 10 levels "(34218.1, 37413.8]",..
 $ MedianAge            : num  33 45 42.8 48.3 45.4 51.7 35.8 54.4 45.2 ...
 $ MedianAgeMale        : num  32.2 44 42.2 47.8 43.5 50.8 34.7 54 44.9 ...
 $ MedianAgeFemale      : num  33.7 45.8 43.4 48.9 48 52.5 37 54.6 45.5 ...
 $ Geography            : Factor w/ 3047 levels "Abbeville County",..
 $ AvgHouseholdSize     : num  2.34 2.62 2.52 2.34 2.58 2.24 2.65 2.04 ...
 $ PercentMarried       : num  44.5 54.2 52.7 57.8 50.4 52.7 50 56.8 54.4 ...
 $ PctNoHS18_24         : num  6.1 24 20.2 14.9 29.9 27.3 15.6 17.7 20 10.9 ...
 $ PctHS18_24           : num  22.4 36.6 41.2 43 35.1 33.9 36.3 32.4 ...
 $ PctBachDeg18_24      : num  7.5 9.5 2.5 2 4.5 2.2 7.1 5.2 2.4 8.6 ...
 $ PctHS25_Over         : num  26 29 31.6 33.4 30.4 31.6 28.8 17.2 29.2 ...
 $ PctBachDeg25_Over    : num  22.7 16 9.3 15 11.9 11.3 16.2 26.2 14.2 18.1 ...
 $ PctEmployed16_Over   : num  55.9 45.9 48.3 48.2 44.1 40.9 56.6 54.6 51.5 ...
 $ PctUnemployed16_Over : num  7.8 7 12.1 4.8 12.9 8.9 9.2 5.9 8.3 8.4 ...
 $ PctPrivateCoverage   : num  70.2 63.7 58.4 61.6 60 55.8 69.9 67.2 64.4 ...
```

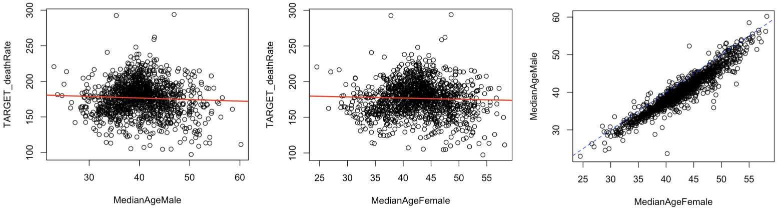
Moindres carrés

```r
with(database,plot(incidenceRate,TARGET_deathRate))
abline(lm(TARGET_deathRate~incidenceRate,data=database),lwd=2,col="red")
with(database,plot(medIncome,TARGET_deathRate))
abline(lm(TARGET_deathRate~medIncome,data=database),lwd=2,col="red")
with(database,plot(MedianAge,TARGET_deathRate))
abline(lm(TARGET_deathRate~MedianAge,data=database),lwd=2,col="red")
```

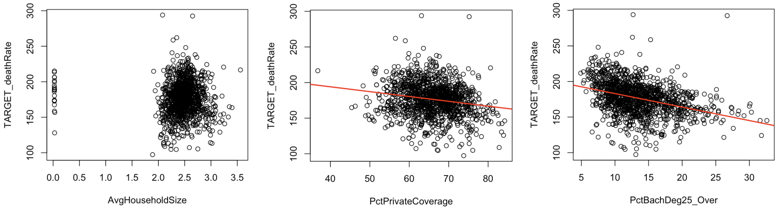



```r
idx =which(database$MedianAge>300)
database = database[-idx,]
with(database,plot(MedianAgeMale,TARGET_deathRate))
abline(lm(TARGET_deathRate~MedianAgeMale,data=database),lwd=2,col="red")
with(database,plot(MedianAgeFemale,TARGET_deathRate))
abline(lm(TARGET_deathRate~MedianAgeFemale,data=database),lwd=2,col="red")
with(database,plot(MedianAgeFemale,MedianAgeMale))
```

Moindres carrés




```r
with(database,plot(AvgHouseholdSize,TARGET_deathRate))
idx = which(database$AvgHouseholdSize<1)
database = database[-idx,]
with(data=database,plot(PctPrivateCoverage,TARGET_deathRate))
abline(lm(TARGET_deathRate~PctPrivateCoverage,data=database),lwd=2,col="red")
with(data=database,plot(PctBachDeg25_Over,TARGET_deathRate))
```

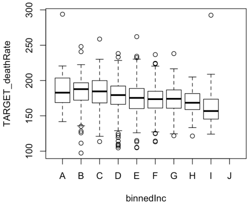


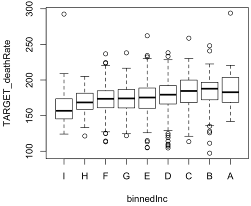
Moindres carrés

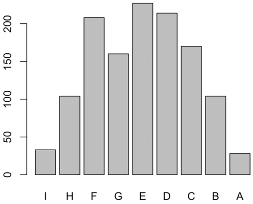
```
> reg_simple = lm(TARGET_deathRate~avgAnnCount+MedianAgeMale+incidenceRate+
     medIncome,data=database)
> summary(reg_simple)

Call:
lm(formula = TARGET_deathRate ~ avgAnnCount + MedianAgeMale +
    incidenceRate + medIncome, data = database)

Residuals:
   Min      1Q  Median      3Q     Max
-67.50  -12.08   -0.60   12.10  130.26

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    1.531e+02  9.609e+00   15.932   < 2e-16 ***
avgAnnCount   -1.024e-03  3.280e-04   -3.122   0.00184 **
MedianAgeMale -5.623e-01  1.167e-01   -4.818  1.63e-06 ***
incidenceRate  2.097e-01  1.228e-02   17.076   < 2e-16 ***
medIncome     -1.006e-03  1.013e-04   -9.932   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.09 on 1243 degrees of freedom
Multiple R-squared:  0.2471, Adjusted R-squared:  0.2446
F-statistic:   102 on 4 and 1243 DF,  p-value: < 2.2e-16
```

Moindres carrés

```
> str(database$Geography)
 Factor w/ 3047 levels "Abbeville County, South Carolina",..: 1460 1464 1589 ...
> str(database$binnedInc)
 Factor w/ 10 levels "(34218.1, 37413.8]",..: 6 6 4 6 7 2 8 8 7 6 ...
> levels(database$binnedInc)=LETTERS[1:10]
with(data = database, boxplot(TARGET_deathRate ~ binnedInc))
A = with(data = database, aggregate(TARGET_deathRate,by=list(binnedInc),FUN=mean))
A = A[order(A$x),]
L = as.character(A$Group.1)
database$binnedInc = factor(database$binnedInc, level=L)
with(data = database, boxplot(TARGET_deathRate ~ binnedInc))
```







```
> reg=lm(TARGET_deathRate ~ binnedInc, data=database)
> summary(reg)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  162.509      3.939  41.253  < 2e-16 ***
binnedIncH     6.524      4.521   1.443 0.149312
binnedIncF    10.340      4.240   2.439 0.014883 *
binnedIncG    11.762      4.327   2.719 0.006647 **
binnedIncE    12.747      4.216   3.024 0.002549 **
binnedIncD    16.015      4.232   3.784 0.000162 ***
binnedIncC    19.943      4.305   4.633 3.99e-06 ***
binnedIncB    21.064      4.521   4.659 3.52e-06 ***
binnedIncA    25.309      5.814   4.353 1.45e-05 ***
> pairwise.t.test(database$TARGET_deathRate,database$binnedInc)

 Pairwise comparisons using t tests with pooled SD

data:  database$TARGET_deathRate and database$binnedInc

  I       H       F       G       E       D       C       B
H 1.00000 -       -       -       -       -       -       -
F 0.23813 1.00000 -       -       -       -       -       -
G 0.11964 0.80875 1.00000 -       -       -       -       -
E 0.05353 0.30537 1.00000 1.00000 -       -       -       -
D 0.00452 0.01258 0.17214 0.80875 1.00000 -       -       -
C 0.00014 7.7e-05 0.00134 0.02717 0.04047 0.91417 -       -
B 0.00012 0.00014 0.00252 0.02717 0.04299 0.80875 1.00000 -
A 0.00047 0.00295 0.02717 0.07077 0.10766 0.57693 1.00000 1.00000

P value adjustment method: holm
```

Moindres carrés

```
> library(car)
> database$binnedInc = relevel(database$binnedInc, "G")
> reg = lm(TARGET_deathRate ~ binnedInc, data=database)
> summary(reg)

            Estimate Std. Error t value Pr(>|t|)
binnedIncI  -11.6648     4.3125  -2.705 0.006924 **
binnedIncH   -5.3206     2.8149  -1.890 0.058963 .
binnedIncF   -0.9641     2.3545  -0.409 0.682262
binnedIncE    1.2690     2.3129   0.549 0.583334
binnedIncD    4.7566     2.3340   2.038 0.041761 *
> linearHypothesis(reg, c("binnedIncF = 0",
+                         "binnedIncE = 0"))

Model 1: restricted model
Model 2: TARGET_deathRate ~ binnedInc

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1275 648943
2   1273 648382  2    560.99 0.5507 0.5767
> linearHypothesis(reg, c("binnedIncH = 0",
+                         "binnedIncF = 0",
+                         "binnedIncE = 0"))

Model 1: restricted model
Model 2: TARGET_deathRate ~ binnedInc

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1276 651663
2   1273 648382  3      3281 2.1473 0.09254 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> levels(database$binnedInc) = c("EFGH","I","EFGH","EFGH","EFGH","D","C","B","A
")
```

Moindres carrés

```
> database$binnedInc = relevel(database$binnedInc, "C")
> reg=lm(TARGET_deathRate ~ binnedInc, data=database)
> summary(reg)

Call:
lm(formula = TARGET_deathRate ~ binnedInc, data = database)

Residuals:
    Min      1Q  Median      3Q     Max
-86.444 -13.185   0.497  13.323 129.991

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    182.368      1.699 107.362  < 2e-16 ***
binnedIncEFGH   -8.865      1.897  -4.673 3.29e-06 ***
binnedIncI     -19.859      4.285  -4.634 3.95e-06 ***
binnedIncD      -3.437      2.275  -1.511    0.131
binnedIncB       1.376      2.784   0.494    0.621
binnedIncA       5.484      4.527   1.211    0.226
> linearHypothesis(reg, c("binnedIncA = 0",
+                         "binnedIncB = 0",
+                         "binnedIncD = 0"))

Model 1: restricted model
Model 2: TARGET_deathRate ~ binnedInc

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   1279 654967
2   1276 651663  3    3304.2 2.1566 0.09141 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> levels(database$binnedInc) = c("ABCD","EFGH","I","ABCD", "ABCD","ABCD")
```

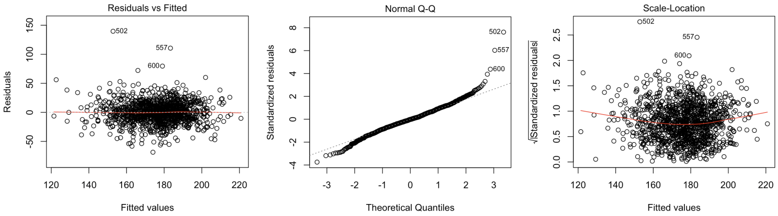
```
> v_initial = lm(TARGET_deathRate~.-Geography ,data=database)
> Backward_regression = step(v_initial , direction = "backward")
Start:  AIC=7225.63

                          Df Sum of Sq    RSS    AIC
- povertyPercent           1        53 383964 7223.8
- avgAnnCount              1        81 383992 7223.9
- binnedInc                8      4414 388325 7223.9
- PctBlack                 1        89 384000 7223.9
- PctUnemployed16_Over     1       100 384011 7224.0
- AvgHouseholdSize         1       193 384104 7224.3
- studyPerCap              1       221 384132 7224.3
<none>                                 383911 7225.6
- medIncome                1       637 384548 7225.7
- avgDeathsPerYear         1       655 384566 7225.8
- popEst2015               1       737 384648 7226.0
- PctPublicCoverageAlone   1       860 384771 7226.4
- PctEmpPrivCoverage       1      1329 385240 7227.9
- PctNoHS18_24             1      1373 385284 7228.1
- PctWhite                 1      1682 385593 7229.1
- PctHS18_24               1      1799 385710 7229.5
- MedianAge                1      1821 385731 7229.5
- MedianAgeMale            1      2037 385948 7230.2
- PctPrivateCoverage       1      2205 386115 7230.8
- PctAsian                 1      2260 386171 7231.0
- MedianAgeFemale          1      2263 386174 7231.0
- PctPublicCoverage        1      2450 386361 7231.6
- PctBachDeg18_24          1      3038 386949 7233.5
- BirthRate                1      4288 388199 7237.5
- PctHS25_Over             1      4507 388418 7238.2
- PctOtherRace             1      4929 388840 7239.5
- PctEmployed16_Over       1      5158 389069 7240.3
- PctBachDeg25_Over        1      5385 389296 7241.0
- PercentMarried           1      6383 390294 7244.2
```

```
Step:  AIC=7212.13
TARGET_deathRate ~ incidenceRate + popEst2015 + MedianAge + MedianAgeMale +
    MedianAgeFemale + PercentMarried + PctNoHS18_24 + PctHS18_24 +
    PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over + PctEmployed16_Over +
    PctPrivateCoverage + PctEmpPrivCoverage + PctPublicCoverage +
    PctPublicCoverageAlone + PctWhite + PctAsian + PctOtherRace +
    PctMarriedHouseholds + BirthRate

                         Df Sum of Sq    RSS    AIC
<none>                                389646 7212.1
- popEst2015              1       678 390324 7212.3
- PctPublicCoverageAlone  1       797 390442 7212.7
- PctHS18_24              1      1571 391217 7215.2
- PctNoHS18_24            1      1680 391326 7215.5
- PctAsian                1      1771 391417 7215.8
- MedianAge               1      2067 391712 7216.7
- MedianAgeMale           1      2178 391823 7217.1
- PctPublicCoverage       1      2178 391824 7217.1
- PctEmpPrivCoverage      1      2203 391848 7217.2
- PctBachDeg18_24         1      2355 392001 7217.7
- MedianAgeFemale         1      2622 392268 7218.5
- PctPrivateCoverage      1      3567 393213 7221.5
- PctWhite                1      4111 393756 7223.2
- PctBachDeg25_Over       1      4356 394001 7224.0
- BirthRate               1      4465 394110 7224.4
- PctHS25_Over            1      5521 395166 7227.7
- PctOtherRace            1      5835 395480 7228.7
- PercentMarried          1      9975 399621 7241.7
- PctEmployed16_Over      1     10177 399822 7242.3
- PctMarriedHouseholds    1     16972 406617 7263.3
- incidenceRate           1     59602 449247 7387.8
```

Moindres carrés

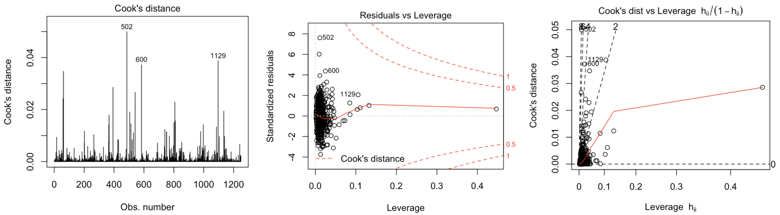
```
> reg_complex = lm(TARGET_deathRate ~ avgDeathsPerYear + incidenceRate +
+                    popEst2015 + MedianAgeMale + PercentMarried + PctHS18_24 +
+                    PctHS25_Over + PctBachDeg25_Over + PctEmployed16_Over +
     PctPublicCoverage +
+                    PctOtherRace + PctMarriedHouseholds + BirthRate,data=
     database)
> summary(reg_complex)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.228e+02  1.606e+01  13.872  < 2e-16 ***
avgDeathsPerYear     5.878e-03  4.449e-03   1.321   0.1867
incidenceRate        1.631e-01  1.203e-02  13.564  < 2e-16 ***
popEst2015          -9.802e-06  6.570e-06  -1.492   0.1360
MedianAgeMale       -1.151e+00  1.814e-01  -6.349 3.05e-10 ***
PercentMarried       1.096e+00  2.548e-01   4.303 1.82e-05 ***
PctHS18_24           3.532e-01  6.777e-02   5.212 2.19e-07 ***
PctHS25_Over         3.443e-01  1.350e-01   2.550   0.0109 *
PctBachDeg25_Over   -1.375e+00  2.199e-01  -6.251 5.62e-10 ***
PctEmployed16_Over  -7.004e-01  1.343e-01  -5.216 2.15e-07 ***
PctPublicCoverage   -2.266e-01  1.569e-01  -1.444   0.1489
PctOtherRace        -4.315e-01  1.795e-01  -2.404   0.0163 *
PctMarriedHouseholds -1.597e+00  2.388e-01  -6.686 3.46e-11 ***
BirthRate           -1.214e+00  2.913e-01  -4.169 3.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.43 on 1234 degrees of freedom
Multiple R-squared:  0.3712, Adjusted R-squared:  0.3646
F-statistic: 56.03 on 13 and 1234 DF,  p-value: < 2.2e-16
```

```
> plot(reg_complex, which=1:3)
```



Avec le nuage de points $(\widehat{y}_i, \widehat{\varepsilon}_i)$ à gauche, un QQ-plot de normalité au centre $\left(\widetilde{\varepsilon}_{i:n}, \Phi^{-1}\left(\frac{i}{n}\right)\right)$, et $(\widehat{Y}_i, \sqrt{|\widetilde{\varepsilon}_i|})$ à droite

```
1 > plot(reg_complex, which=4:6)
```



Rappelons que la distance de Cook est

$$C_i = \frac{\widehat{\varepsilon}_i^2}{p \cdot \text{MSE}} \cdot \left(\frac{H_{i,i}}{(1-H_{i,i})^2}\right) \text{ avec } \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}} = [H_{i,i}],$$

où le terme $H_{i,i}$ est le *leverage*, et les résidus Studentisés sont

$$\widehat{r}_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{1-H_{i,i}}}$$

On a au centre $(H_{i,i}, \widehat{r}_i)$ et à droite $(H_{i,i}, C_i)$

Moindres carrés

```
> which(cooks.distance(reg3_partie2)>.03)
  65  408  502  600 1129
> B=database[,c("TARGET_deathRate","avgDeathsPerYear","incidenceRate","popEst2015","MedianAgeMale",
+ "PercentMarried","PctHS18_24","PctHS25_Over","PctBachDeg25_Over","PctEmployed16_Over",
+ "PctPublicCoverage","PctOtherRace","PctMarriedHouseholds","BirthRate")]
> q1=apply(B,2,function(x) quantile(x,.1))
> q9=apply(B,2,function(x) quantile(x,.9))
> m =apply(B,2,mean)
> cbind(Q1=q1,M=m,Q9=q9,t(B[which(cooks.distance(reg3_partie2)>.03),]))
                   Q1        M       Q9     65        408      502    600    1129
TARGET_deathRt  148.2    176.6    205.4  121.8      148.4    292.5  258.7   220.6
avgDeathsPerYr   14.0    227.1    472.9    9.0    14010.0    269.0   10.0     9.0
incidenceRate   392.1    450.1    505.4  453.5      405.5    460.5  456.9   510.8
popEst2015     5550.7 125856.9 252979.7 6634.0 10170292.0 103465.0 2216.0 11368.0
MedianAgeMale    33.8     40.2     47.0   39.5       34.4     35.4   42.9    23.0
PercentMarried   46.2     52.7     59.4   44.8       42.4     52.3   60.9    46.8
PctHS18_24       23.8     35.1     46.1   28.9       27.0     22.5   44.4    40.0
PctHS25_Over     26.7     35.2     43.8   35.4       20.7     16.0   36.3    27.0
PctBachDeg25_Ov   8.7     13.5     19.5    7.5       19.8     26.7   15.3    19.3
PctEmployed16_O  47.5     55.1     62.5   36.9       58.0     62.9   60.8    24.0
PctPublicCovrg   28.6     36.0     43.4   34.7       32.9     26.6   38.4    16.1
PctOtherRace      0.1      2.0      5.0    2.9       19.6      1.9   12.2     2.9
PctMarriedHlds   45.8     51.8     57.6   51.4       44.6     51.2   52.4    67.3
BirthRate         3.6      5.6      7.8   14.6        4.7      4.8    2.2    11.7
```

```
> library(leaps)
> forward = regsubsets(TARGET_deathRate ~.-Geography,data = database, method = "
      forward", nbest=1)
backward = regsubsets(TARGET_deathRate ~.-Geography,data = database, method = "
      backward", nbest=1)
stepwise = regsubsets(TARGET_deathRate ~.-Geography, data = database, method = "
      seqrep", nbest=1)
> best_subset = regsubsets(TARGET_deathRate ~.-Geography,data = database, method
       = "exhaustive", nbest=1)
> plot(best_subset, scale = "adjr2", main = "Forward Selection")
```