Exercice Modèle Linéaire et ANOVA

18 décembre 2023

Le jour du dépassement (overshoot day en anglais) correspond à la date de l'année, calculée par l'ONG américaine Global Footprint Network, à partir de laquelle l'humanité est supposée avoir consommé l'ensemble des ressources naturelles que la planète est capable de produire en un an pour régénérer ses consommations ou absorber les déchets produits, dont le dioxyde de carbone. Passé cette date, l'humanité puiserait donc dans ses ressources à une vitesse qui n'est pas de l'ordre du « renouvelable à échelle humaine », accumulant les déchets au-delà de leur absorption sur le reste de l'année en cours.

Cette mesure peut être faite à l'échelle mondiale mais également à l'échelle de chaque pays, où la date de dépassement est calculée proportionnellement aux ressources allouées à chaque pays en fonction de différents critères, en particulier sa taille et son nombre d'habitants. Le jeu de données contient notamment les informations suivantes (par pays) :

- l'espérance de vie
- l'indice de développement humain (hdi)
- le PIB par habitant (per_capita_gdp)
- la région du monde
- la population (en million d'habitants)
- le jour du dépassement

NB : La base de données complète est le fichier "NFA 2022 Public Data Package 1.1.xlsx", téléchargée sur https://www.footprintnetwork.org/ . Ici, on a utilisé un pré-traitement des données grâce au code contenu et expliqué dans le fichier Preprocessing.Rmd

Influence des variables quantitatives

On s'intéresse à l'influence des variables quantitatives, à savoir l'indice de développement humain (hdi), le PIB (per_capita_gdp), la taille de la population (pop) et l'espérance de vie (life_expectancy), sur le jour de dépassement. On souhaite notamment répondre aux questions suivantes :

- Quelles sont les variables avec un impact sur le jour du dépassement ?
- Quel sous-ensemble de variables choisiriez-vous pour expliquer au mieux le jour du dépassement ?
- Avec le modèle choisi, donnez la valeur prédite pour un pays qui aurait les caractéristiques suivantes : un hdi de 0.8, un PIB de 35000, une espérance de vie de 78 ans et un population de 5 millions d'habitants.

Pour chaque test vous interpréterez les sorties, vous écrirez les hypothèses testées, la sortie R qui vous donne le résultat ainsi que la conclusion.

```
load("Overshoot_day_by_country.Rdata")
overshoot_country=na.omit(overshoot_country)
library(car)
```

```
## Le chargement a nécessité le package : carData
```

```
res=lm(overshoot_day~hdi+per_capita_gdp+pop+life_expectancy,data=overshoot_country)
vif(res)
```

```
##             hdi  per_capita_gdp             pop life_expectancy
##        7.365635        2.009717        1.005830        6.233923
```

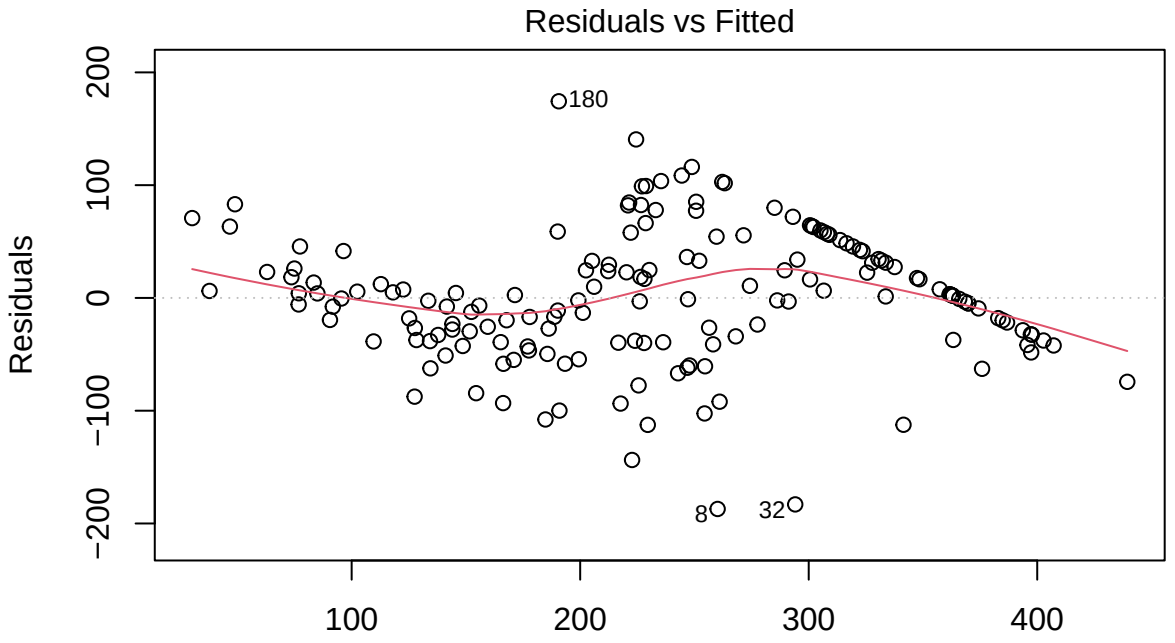
```
summary(res)
```

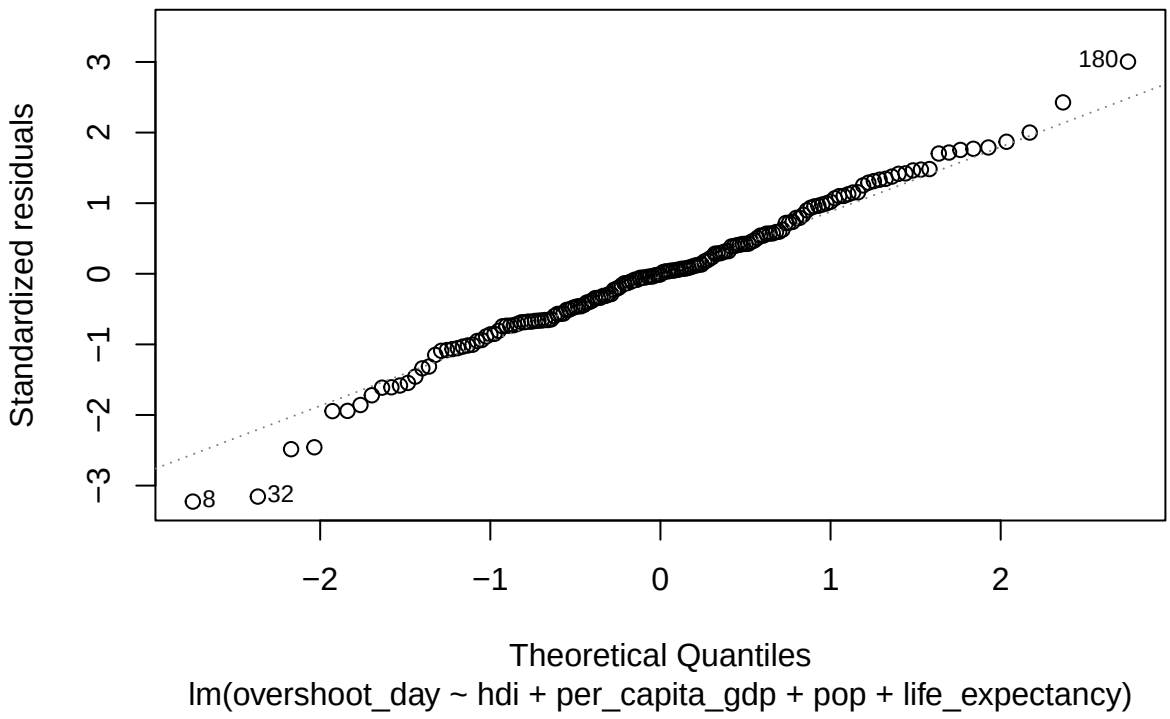
```
##
## Call:
## lm(formula = overshoot_day ~ hdi + per_capita_gdp + pop + life_expectancy,
##     data = overshoot_country)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -187.140  -37.549   -0.516   33.569  174.359
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.337e+02  6.571e+01   6.600 5.58e-10 ***
## hdi             -7.214e+02  8.155e+01  -8.845 1.49e-15 ***
## per_capita_gdp  -1.073e-03  3.250e-04  -3.302  0.00118 **
## pop              1.172e-02  2.873e-02   0.408  0.68375
## life_expectancy  4.645e+00  1.511e+00   3.073  0.00249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.37 on 162 degrees of freedom
## Multiple R-squared:  0.7279, Adjusted R-squared:  0.7212
## F-statistic: 108.4 on 4 and 162 DF,  p-value: < 2.2e-16
```

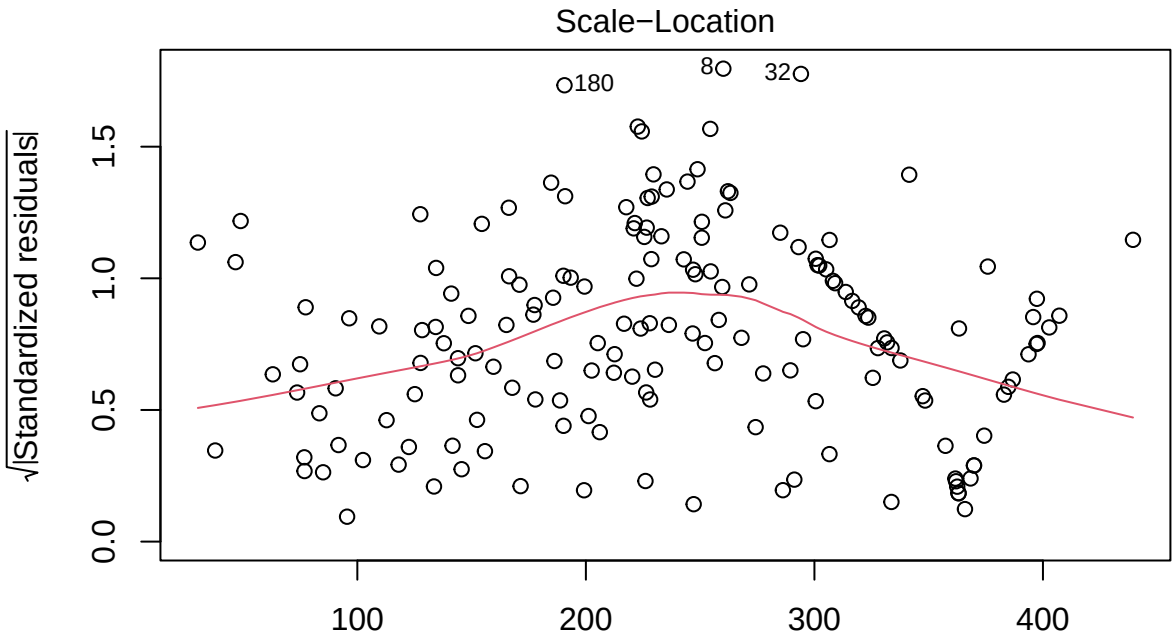
```
anova(res)
```

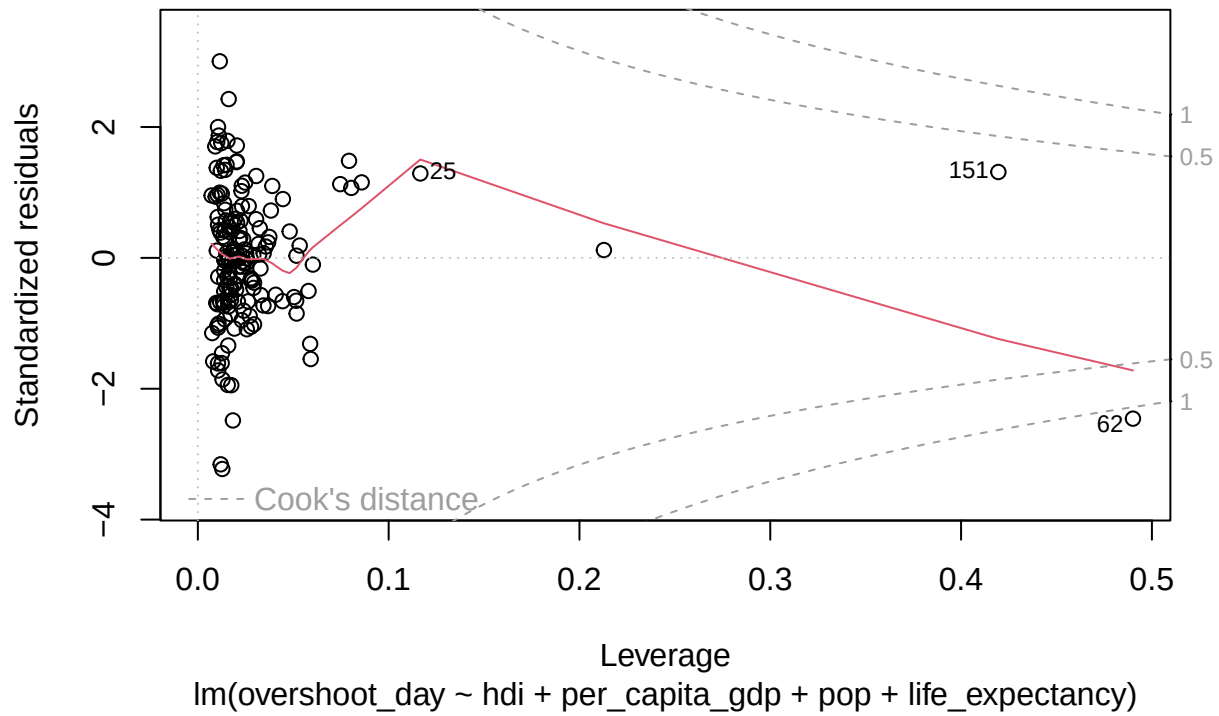
```
## Analysis of Variance Table
##
## Response: overshoot_day
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## hdi                1 1403862 1403862 411.9910 < 2.2e-16 ***
## per_capita_gdp     1   39886   39886  11.7053 0.0007889 ***
## pop                1    1143    1143   0.3356 0.5631950
## life_expectancy    1   32183   32183   9.4447 0.0024851 **
## Residuals        162  552016    3408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(res)
```


Residuals vs Fitted
Residuals
200
100
0
−100
−200
180
8
32
100
200
300
400
Fitted values
lm(overshoot_day ~ hdi + per_capita_gdp + pop + life_expectancy)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
−3
180
8
32
−2
−1
0
1
2
Theoretical Quantiles
lm(overshoot_day ~ hdi + per_capita_gdp + pop + life_expectancy)






```
shapiro.test(res$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res$residuals
```

```
## W = 0.98959, p-value = 0.2589
```

```
t.test(res$residuals,mu=0)
```

```
##
##  One Sample t-test
##
## data:  res$residuals
## t = 6.1774e-16, df = 166, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -8.810275  8.810275
## sample estimates:
##   mean of x
## 2.75659e-15
```

```
res2=lm(overshoot_day~hdi,data=overshoot_country)
res3=lm(overshoot_day~hdi+per_capita_gdp,data=overshoot_country)
res4=lm(overshoot_day~hdi+per_capita_gdp+life_expectancy,data=overshoot_country)
anova(res2,res)
```

```
## Analysis of Variance Table
##
## Model 1: overshoot_day ~ hdi
## Model 2: overshoot_day ~ hdi + per_capita_gdp + pop + life_expectancy
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    165 625229
## 2    162 552016  3     73212 7.1619 0.0001516 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(res3,res)
```

```
## Analysis of Variance Table
##
## Model 1: overshoot_day ~ hdi + per_capita_gdp
## Model 2: overshoot_day ~ hdi + per_capita_gdp + pop + life_expectancy
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    164 585343
## 2    162 552016  2     33327 4.8902 0.008667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(res4,res)
```

```
## Analysis of Variance Table
##
## Model 1: overshoot_day ~ hdi + per_capita_gdp + life_expectancy
## Model 2: overshoot_day ~ hdi + per_capita_gdp + pop + life_expectancy
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    163 552584
## 2    162 552016  1    567.48 0.1665 0.6837
```

Influence d'une variable qualitative

On s'intéresse maintenant à l'effet de la variable région sur le jour du dépassement. Commentez et interprétez les sorties R suivantes, en précisant la différence entre les deux sorties **res_bis** et **res_ter**.

```r
res_bis=lm(overshoot_day~region,data=overshoot_country)
summary(res_bis)
```

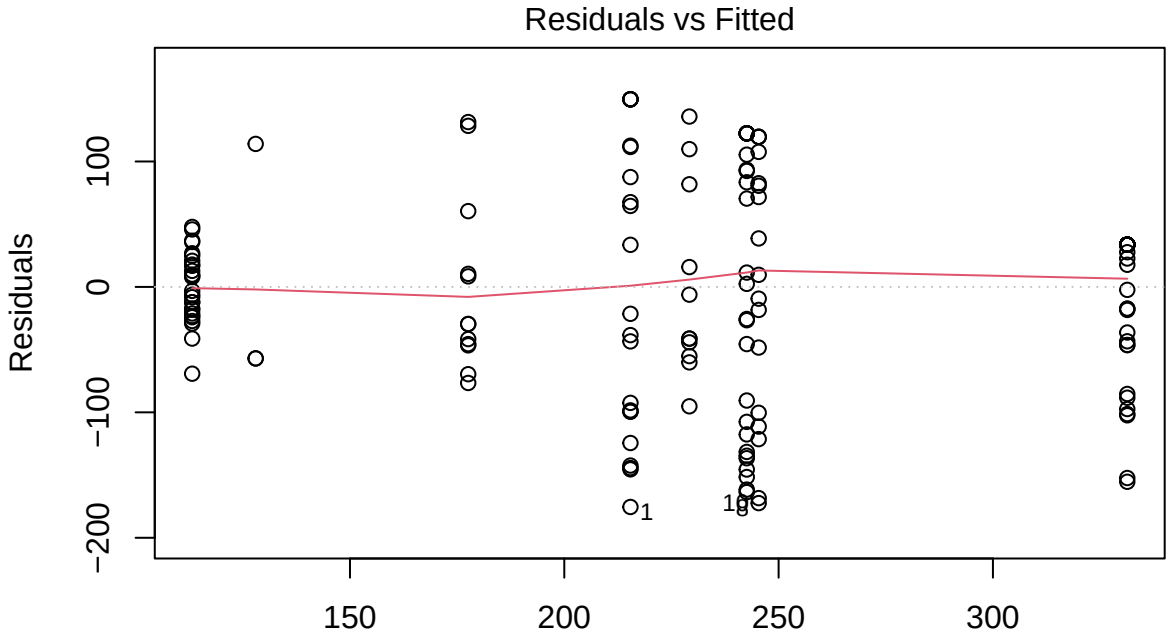
```
##
## Call:
## lm(formula = overshoot_day ~ region, data = overshoot_country)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -175.435  -46.348    9.815   34.733  149.565
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       331.35      12.35  26.837  < 2e-16 ***
## regionAsia-Pacific                -88.78      20.07  -4.423 1.80e-05 ***
## regionCentral America/Caribbean   -85.99      23.77  -3.618 0.000398 ***
## regionEU                         -218.16      20.30 -10.746  < 2e-16 ***
## regionMiddle East/Central Asia   -115.91      21.38  -5.420 2.17e-07 ***
## regionNorth America              -203.35      49.90  -4.075 7.24e-05 ***
## regionOther Europe               -153.76      27.14  -5.665 6.73e-08 ***
## regionSouth America              -102.17      28.11  -3.635 0.000375 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.74 on 159 degrees of freedom
## Multiple R-squared:  0.4505, Adjusted R-squared:  0.4263
## F-statistic: 18.62 on 7 and 159 DF,  p-value: < 2.2e-16
```

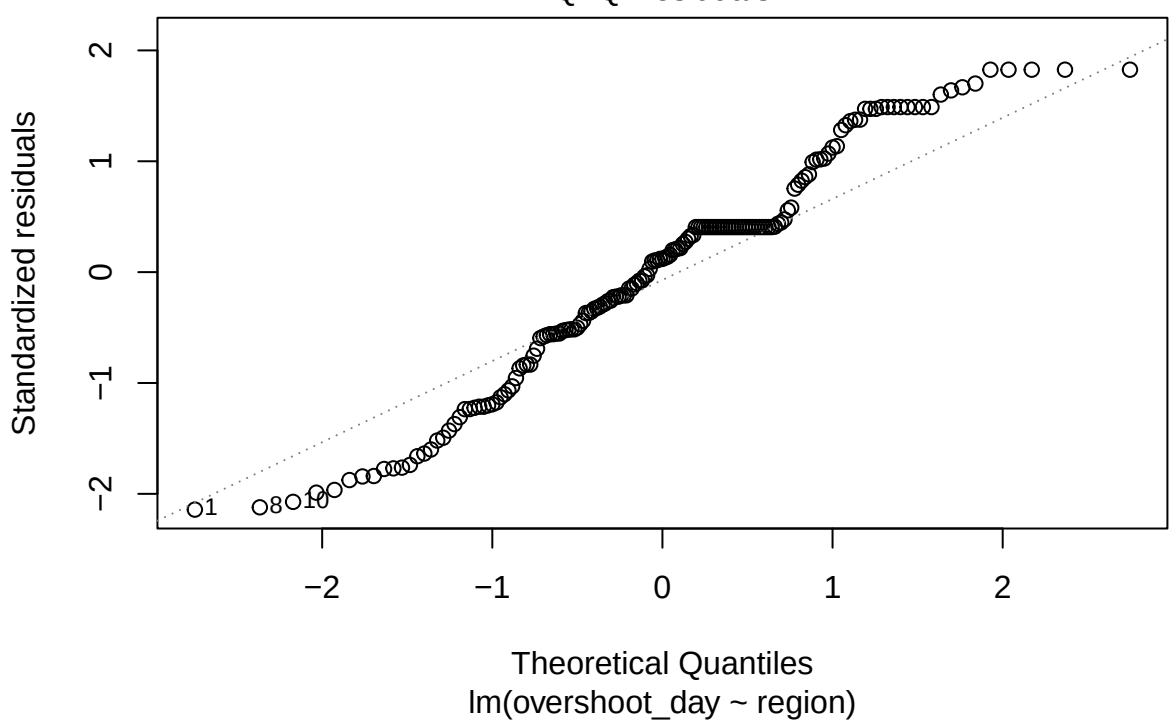
```r
overshoot_country$region=as.factor(overshoot_country$region)
overshoot_country$region=relevel(overshoot_country$region, ref="Asia-Pacific")
res_ter=lm(overshoot_day~region,data=overshoot_country)
summary(res_ter)
```

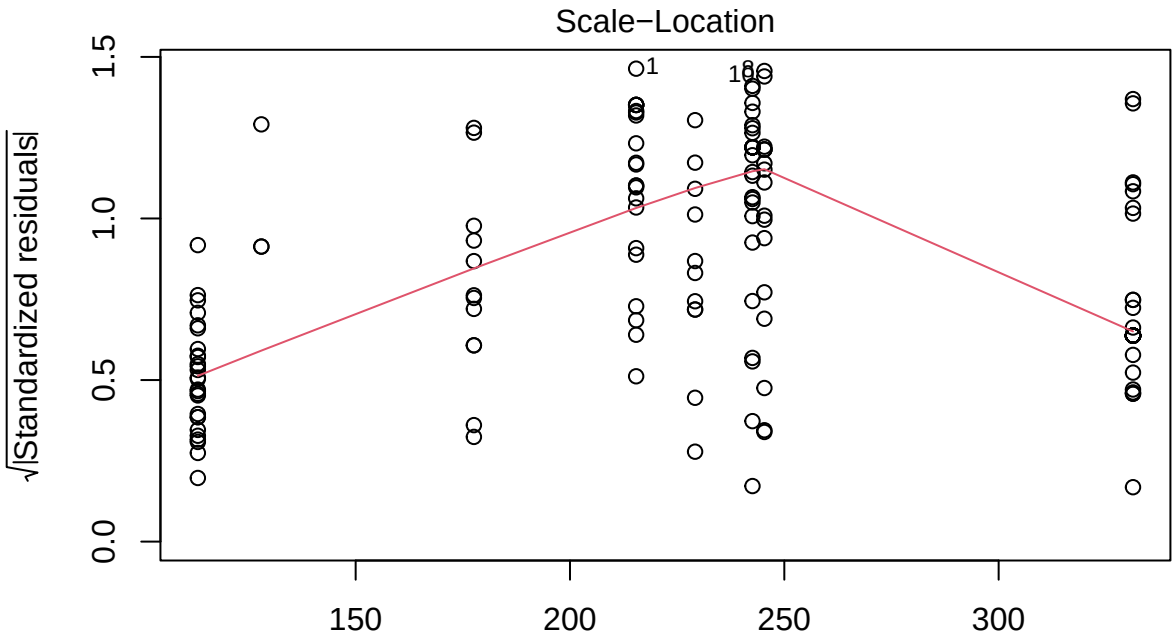
```
##
## Call:
## lm(formula = overshoot_day ~ region, data = overshoot_country)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -175.435  -46.348    9.815   34.733  149.565
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      242.571     15.825  15.328  < 2e-16 ***
## regionAfrica                      88.776     20.072   4.423 1.80e-05 ***
## regionCentral America/Caribbean    2.782     25.747   0.108   0.9141
## regionEU                        -129.386     22.586  -5.729 4.94e-08 ***
## regionMiddle East/Central Asia   -27.137     23.565  -1.152   0.2512
## regionNorth America              -114.571     50.870  -2.252   0.0257 *
## regionOther Europe                -64.988     28.892  -2.249   0.0259 *
## regionSouth America               -13.390     29.797  -0.449   0.6538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

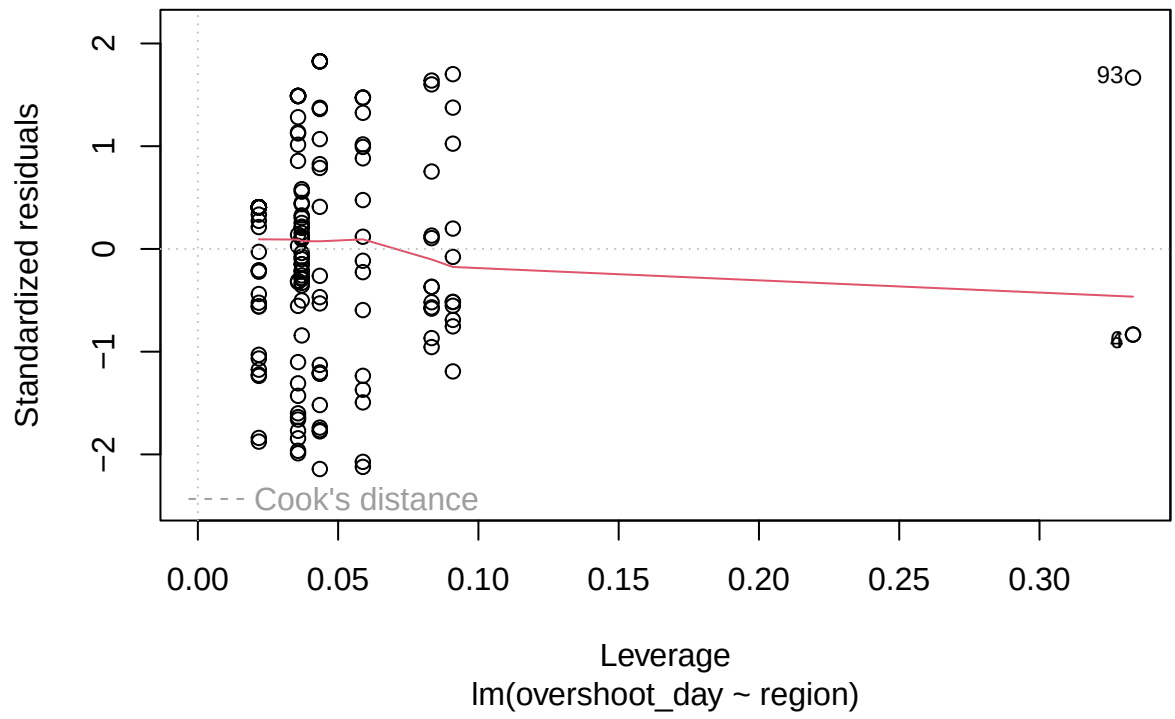
```
## Residual standard error: 83.74 on 159 degrees of freedom
## Multiple R-squared:  0.4505, Adjusted R-squared:  0.4263
## F-statistic: 18.62 on 7 and 159 DF,  p-value: < 2.2e-16
```

```
plot(res_ter)
```






Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
150
200
250
300
Fitted values
lm(overshoot_day ~ region)


Residuals vs Leverage
Standardized residuals
−2
−1
0
1
2
93
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(overshoot_day ~ region)