# model_mida_compressors.R

jcortes

2023-11-30

```
##############################################################################
# Script per estudi tipus de models
# PE 2023-24
# 1. Estimar mitjana
# 2. Estimar diferencia de mitjanes aparellades
# 3. Estimar mitjanes segons categories de factors
# 4. Model lineal simple
# 5. Model lineal multiple
##############################################################################

# rm(list=ls) # Esborrar objectes en memoria

##############################################################################
# Carregar llibreries
# mida: mida original del fitxer
# tar:  mida del fitxer despres de comprimir amb tar
# zip:  mida del fitxer despres de comprimir amb zip
# type: tipus de fitxer
##############################################################################
# install.packages("emmeans")
# install.packages("PairedData")
library(emmeans)
library(PairedData)
```

```
## Loading required package: MASS
```

```
## Loading required package: gld
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
##
## Attaching package: 'PairedData'
```

```
## The following object is masked from 'package:base':
##
##     summary
```

```
################################################################################
# Llegir les dades
# mida: mida original del fitxer
# tar:  mida del fitxer despres de comprimir amb tar
# zip:  mida del fitxer despres de comprimir amb zip
# type: tipus de fitxer
################################################################################
d <- read.csv('https://raw.githubusercontent.com/jordicortes40/PE_Bloc_D/main/Dades/mida_compressors.csv
# d <- read.table('../Dades/mida_compressors.csv',sep=',',header=TRUE, stringsAsFactors = TRUE)
summary(d)
```

```
##       mida              tar               zip              type
##  Min.   :  21504   Min.   :   3157   Min.   :   3228   Length:120
##  1st Qu.: 122880   1st Qu.:  50391   1st Qu.:  50470   Class :character
##  Median : 313291   Median : 199342   Median : 199466   Mode  :character
##  Mean   : 581152   Mean   : 473558   Mean   : 473597
##  3rd Qu.: 784670   3rd Qu.: 647987   3rd Qu.: 648056
##  Max.   :2723867   Max.   :2650696   Max.   :2650683
```

```
d[,1:3] <- d[,1:3]/1000 # Per tenir numeros mes petits

################################################################################
# Model per estimar una mitjana
################################################################################
```

```
mod_1 <- lm(zip~1,data = d)
s <- summary(mod_1)
s
```

```
##
## Call:
## lm(formula = zip ~ 1, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -470.4 -423.1 -274.1  174.5 2177.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   473.60      53.13   8.914 6.93e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 582 on 119 degrees of freedom
```

```
m  <- mod_1$coefficients            # mitjana
se <- s$coefficients[,'Std. Error'] # error estandard
q  <- qt(0.975,mod_1$df)            # quantil
m - q * se
```

```
## (Intercept)
##    368.3935
```

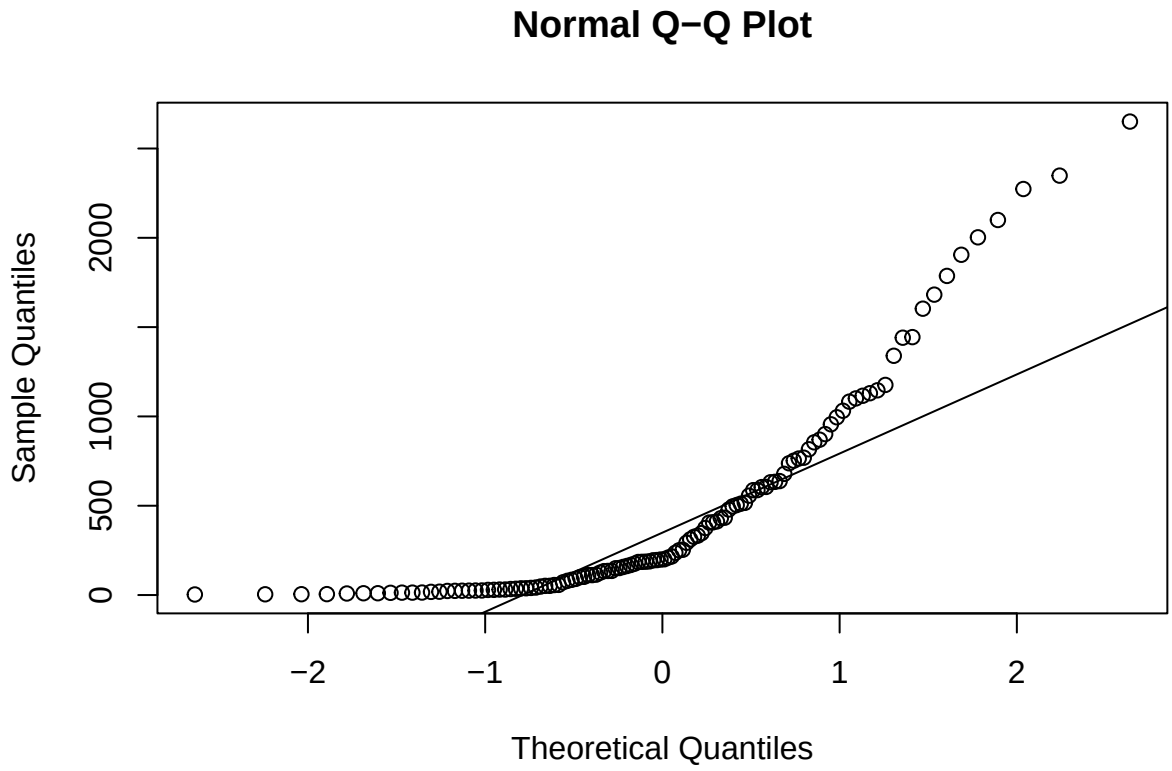
```
m + q * se
```

```
## (Intercept)
##    578.8004
```

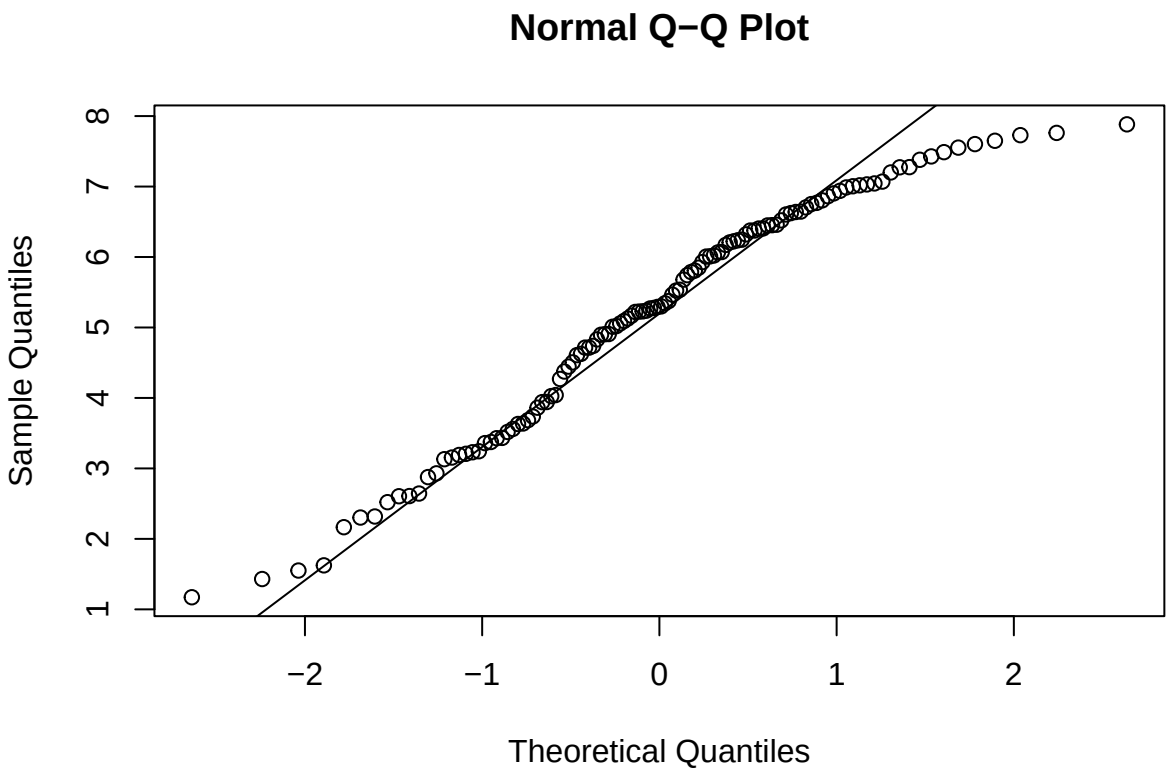
```
confint(mod_1)
```

```
##                2.5 %   97.5 %
## (Intercept) 368.3935 578.8004
```

```
qqnorm(d$zip)
qqline(d$zip)
```




```
qqnorm(log(d$zip))
qqline(log(d$zip))
```




```
##############################################################################
# Model per estimar una diferencia de mitjanes aparellada
##############################################################################
```

```
d$dif <- d$zip-d$tar
```

```
mod_2 <- lm(dif~1,data = d)
s <- summary(mod_2)
s
```

```
##
## Call:
## lm(formula = dif ~ 1, data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.127117 -0.025117 -0.003117  0.025883  0.095883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.039117   0.003416   11.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03742 on 119 degrees of freedom
```

```r
m  <- mod_2$coefficients            # mitjana
se <- s$coefficients[,'Std. Error'] # error estandard
q  <- qt(0.975,mod_1$df)            # quantil
m - q * se
```

```
## (Intercept)
##  0.03235223
```

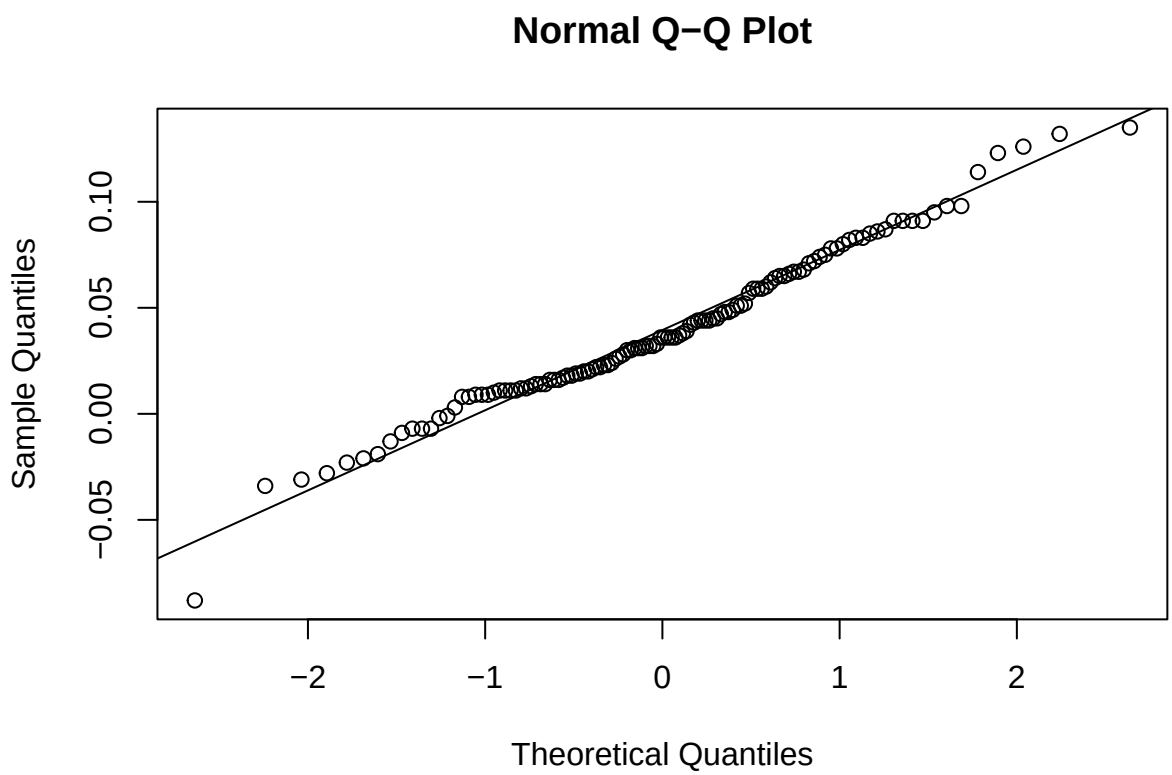
```r
m + q * se
```

```
## (Intercept)
##   0.0458811
```

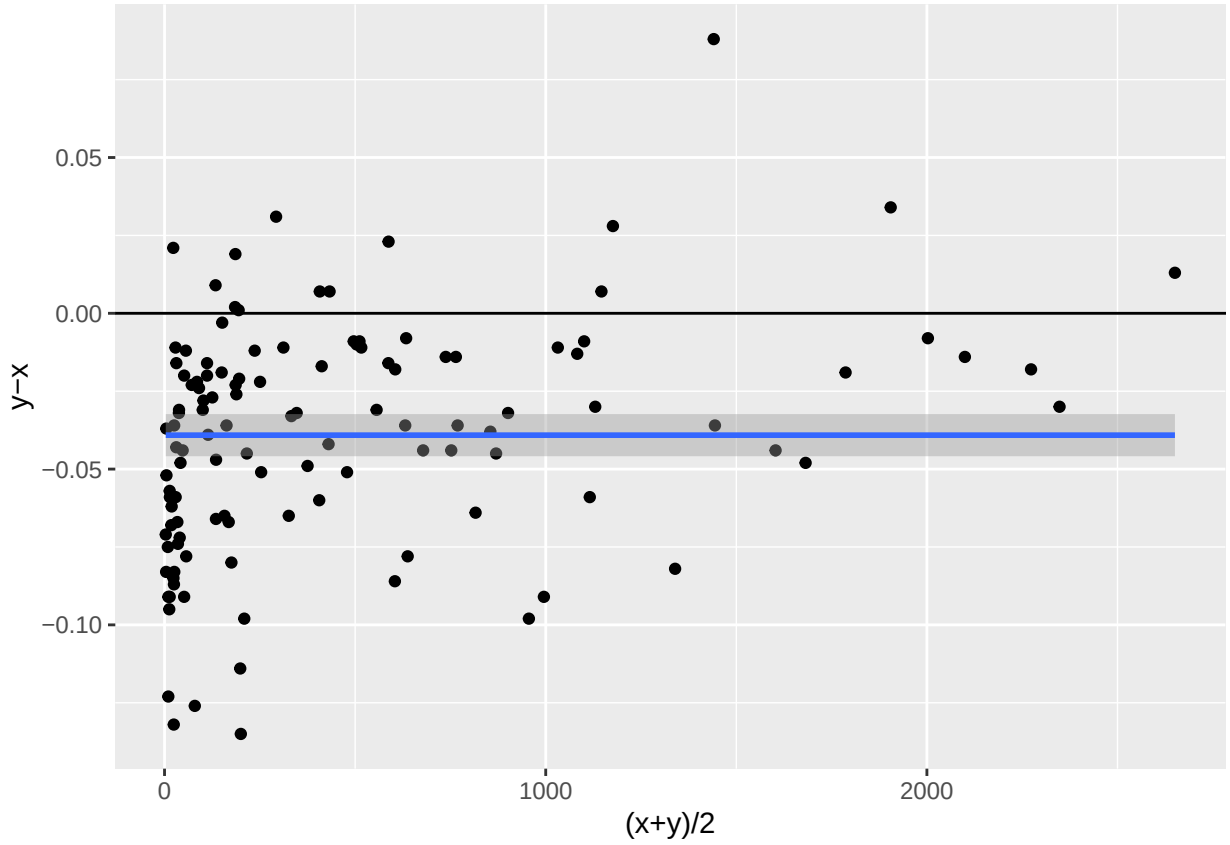
```r
confint(mod_2)
```

```
##                   2.5 %    97.5 %
## (Intercept) 0.03235223 0.0458811
```

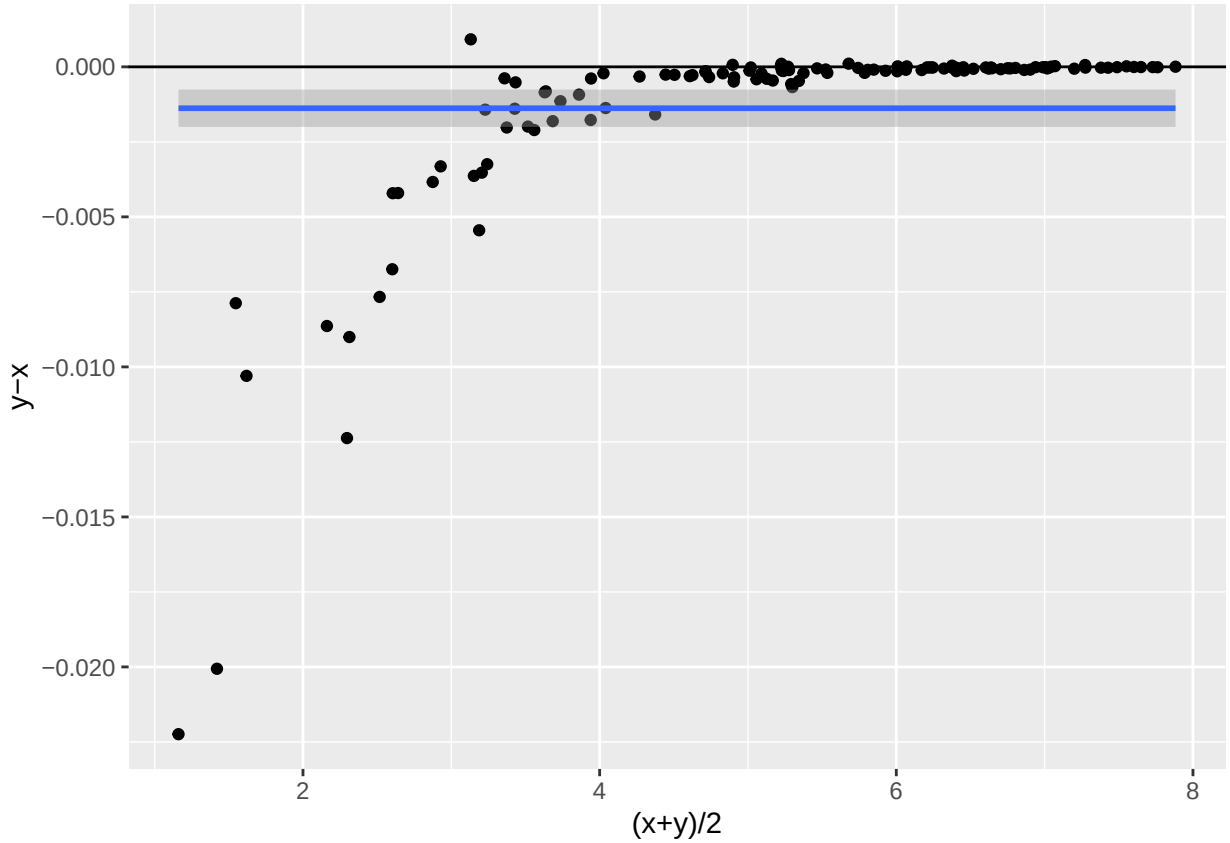
```r
qqnorm(d$dif)
qqline(d$dif)
```



```r
# Sense logs
x <- d$zip
y <- d$tar
p <- paired(x,y)
plot(p,type="BA")
```




```r
# Amb logs --> Empitjora
x <- log(d$zip)
y <- log(d$tar)
p <- paired(x,y)
plot(p,type="BA")
```




```
################################################################################
# Model amb un factor
################################################################################
```

```
d$rati <- d$zip/d$mida
summary(d$rati)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.08184 0.35245 0.79173 0.68470 0.99588 1.00121
```

```
table(d$type)
```

```
##
## doc jpg pdf png ppt xls
##  19  23  22  16  18  22
```

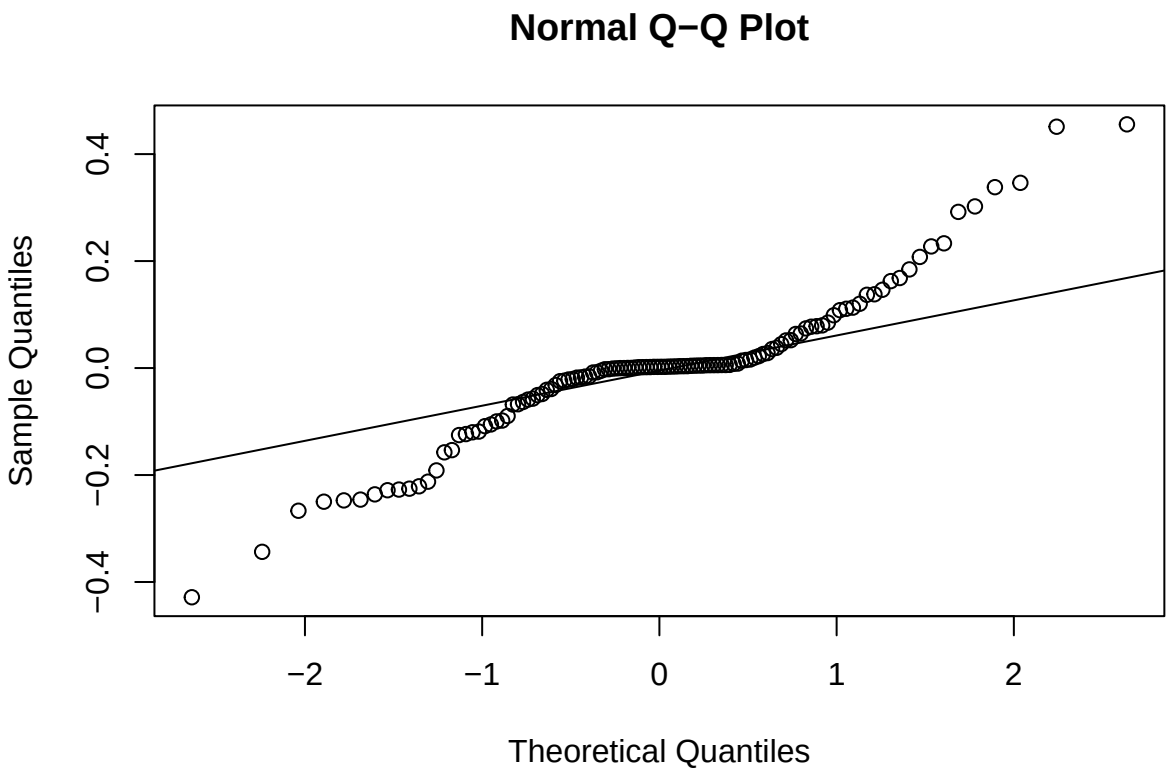
```
with(d,tapply(rati,type,mean))
```

```
##       doc       jpg       pdf       png       ppt       xls
## 0.3712423 0.9967862 0.8460409 0.9948058 0.5802035 0.3277629
```

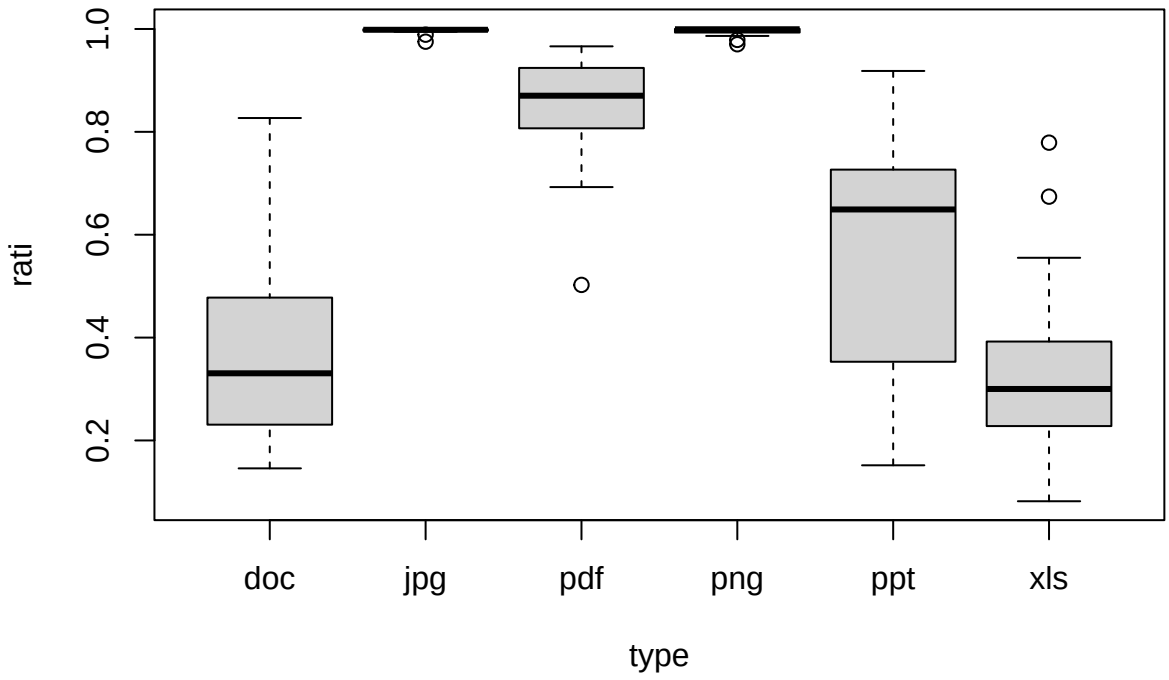
```r
mod_3 <- lm(rati~type,data = d)
s <- summary(mod_3)
s
```

```
##
## Call:
## lm(formula = rati ~ type, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42850 -0.04904  0.00224  0.03953  0.45561
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.37124    0.03274  11.338   <2e-16 ***
## typejpg      0.62554    0.04425  14.138   <2e-16 ***
## typepdf      0.47480    0.04470  10.622   <2e-16 ***
## typepng      0.62356    0.04843  12.877   <2e-16 ***
## typeppt      0.20896    0.04694   4.451    2e-05 ***
## typexls     -0.04348    0.04470  -0.973    0.333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1427 on 114 degrees of freedom
## Multiple R-squared:  0.7988, Adjusted R-squared:   0.79
## F-statistic: 90.51 on 5 and 114 DF,  p-value: < 2.2e-16
```

```r
qqnorm(resid(mod_3))
qqline(resid(mod_3))
```




```
boxplot(rati~type,d)
```



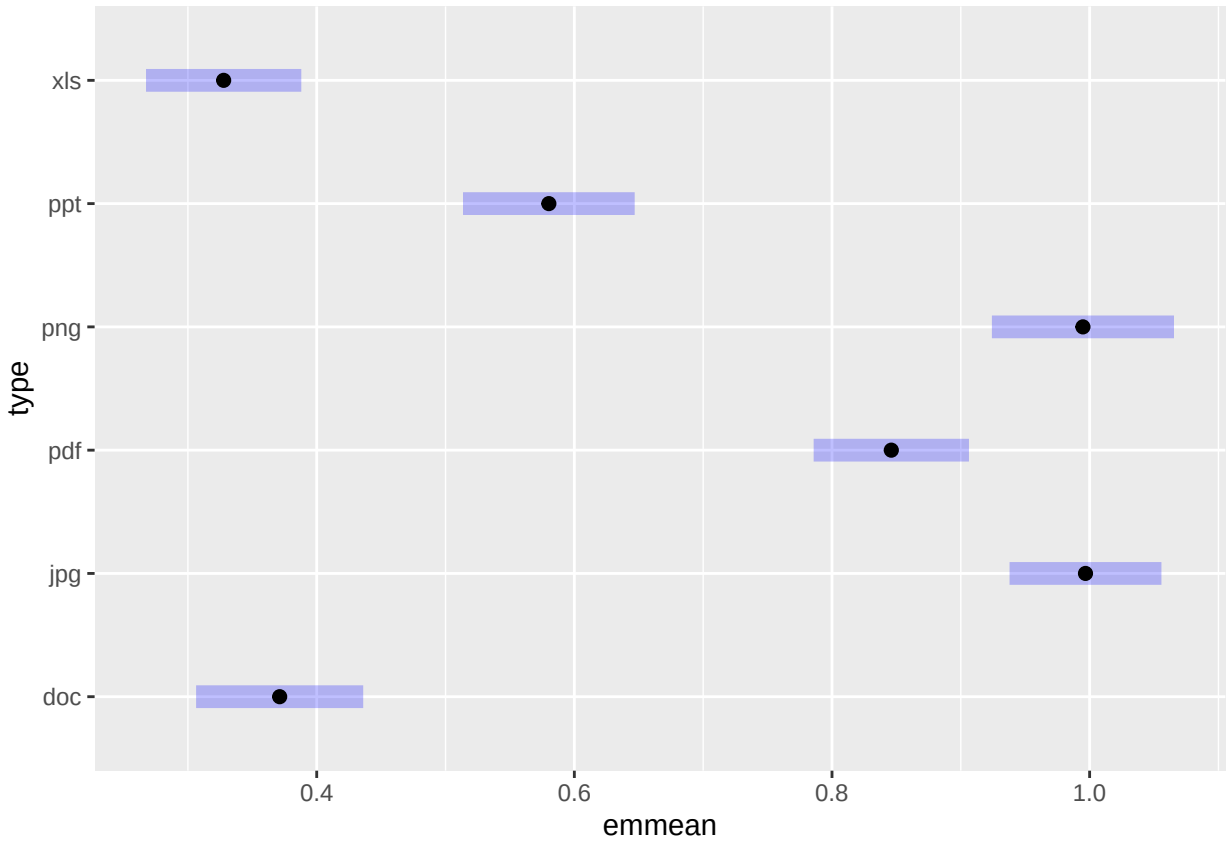

```
confint(mod_3)
```

```
##                 2.5 %     97.5 %
## (Intercept)  0.3063809 0.43610379
## typejpg      0.5378948 0.71319297
## typepdf      0.3862529 0.56334421
## typepng      0.5276321 0.71949478
## typeppt      0.1159680 0.30195440
## typexls     -0.1320250 0.04506623
```

```
em <- emmeans(mod_3,~type)
em
```

```
##  type emmean     SE  df lower.CL upper.CL
##  doc   0.371 0.0327 114    0.306    0.436
##  jpg   0.997 0.0298 114    0.938    1.056
##  pdf   0.846 0.0304 114    0.786    0.906
##  png   0.995 0.0357 114    0.924    1.065
##  ppt   0.580 0.0336 114    0.514    0.647
##  xls   0.328 0.0304 114    0.267    0.388
##
## Confidence level used: 0.95
```

```r
plot(em)   # IC95% graficament
```

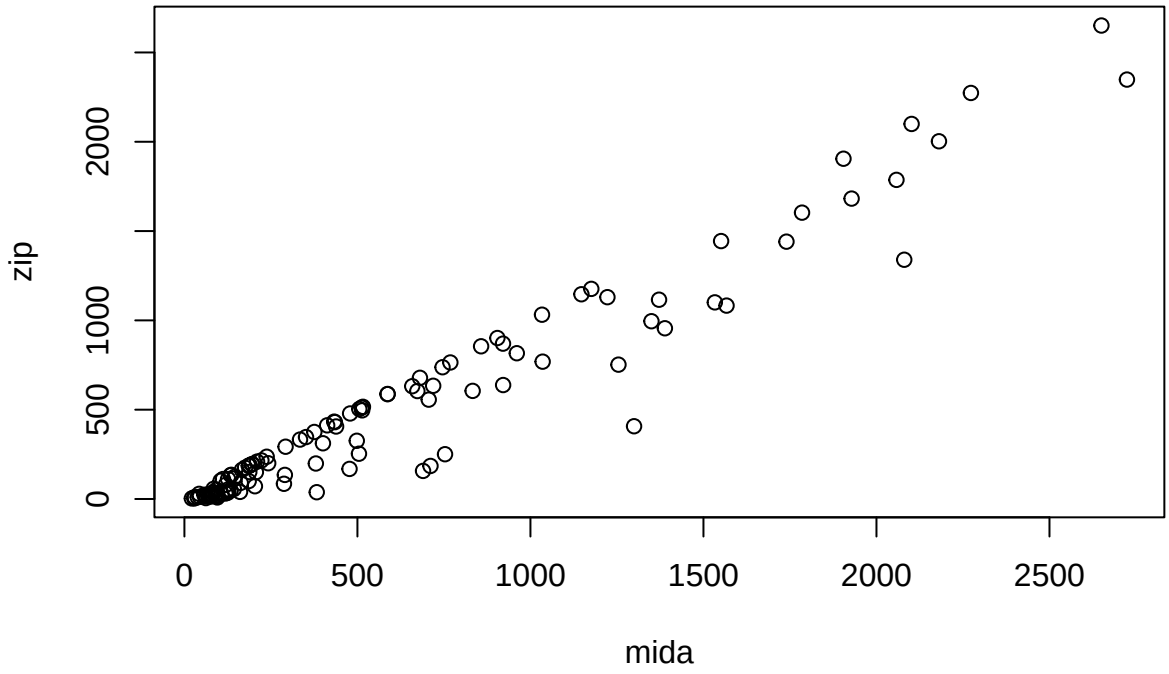



```r
pairs(em)  # Diferencies de mitjanes
```

```
##  contrast  estimate     SE  df t.ratio p.value
##  doc - jpg -0.62554 0.0442 114 -14.138  <.0001
##  doc - pdf -0.47480 0.0447 114 -10.622  <.0001
##  doc - png -0.62356 0.0484 114 -12.877  <.0001
##  doc - ppt -0.20896 0.0469 114  -4.451  0.0003
##  doc - xls  0.04348 0.0447 114   0.973  0.9257
##  jpg - pdf  0.15075 0.0426 114   3.542  0.0074
##  jpg - png  0.00198 0.0465 114   0.043  1.0000
##  jpg - ppt  0.41658 0.0449 114   9.275  <.0001
##  jpg - xls  0.66902 0.0426 114  15.719  <.0001
##  pdf - png -0.14876 0.0469 114  -3.172  0.0233
##  pdf - ppt  0.26584 0.0454 114   5.861  <.0001
##  pdf - xls  0.51828 0.0430 114  12.044  <.0001
##  png - ppt  0.41460 0.0490 114   8.455  <.0001
##  png - xls  0.66704 0.0469 114  14.225  <.0001
##  ppt - xls  0.25244 0.0454 114   5.565  <.0001
##
## P value adjustment: tukey method for comparing a family of 6 estimates
```

```
##############################################################################
# Model lineal simple (MLS)
##############################################################################
```

```
plot(zip~mida,data = d)
```

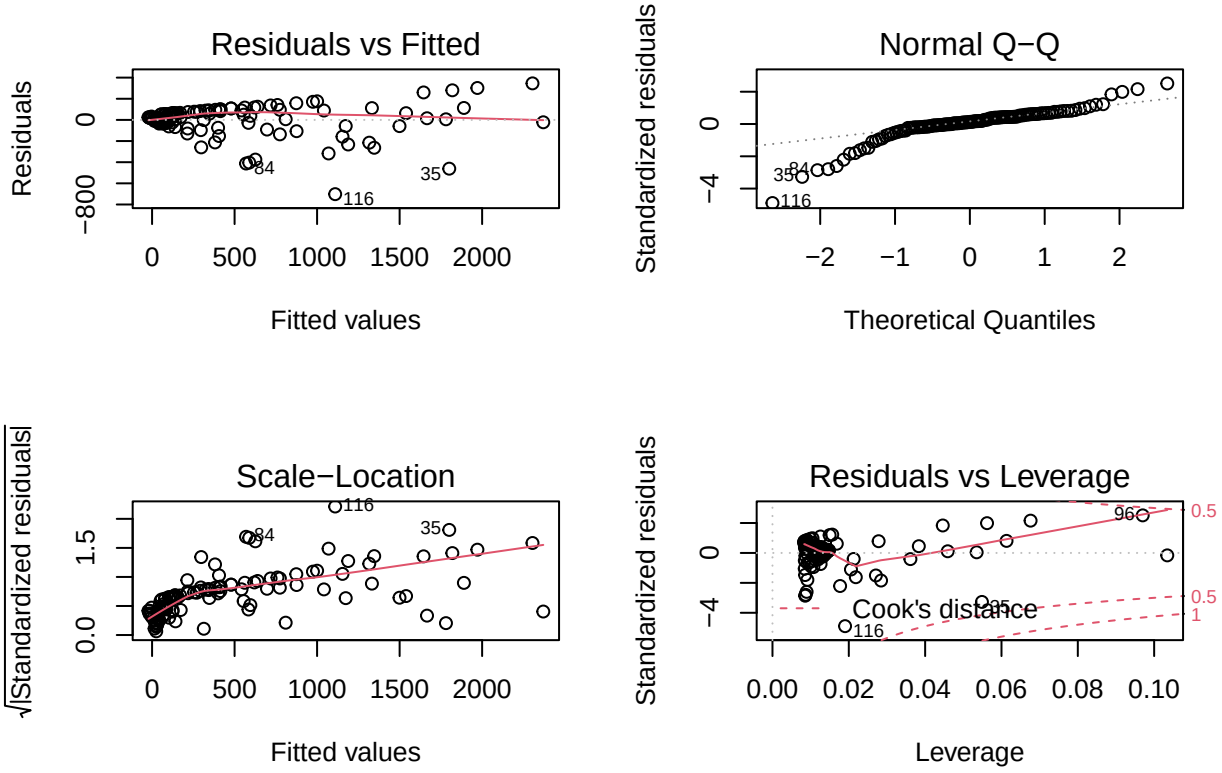



```
mod_4 <- lm(zip~mida,data = d)
s <- summary(mod_4)
s
```

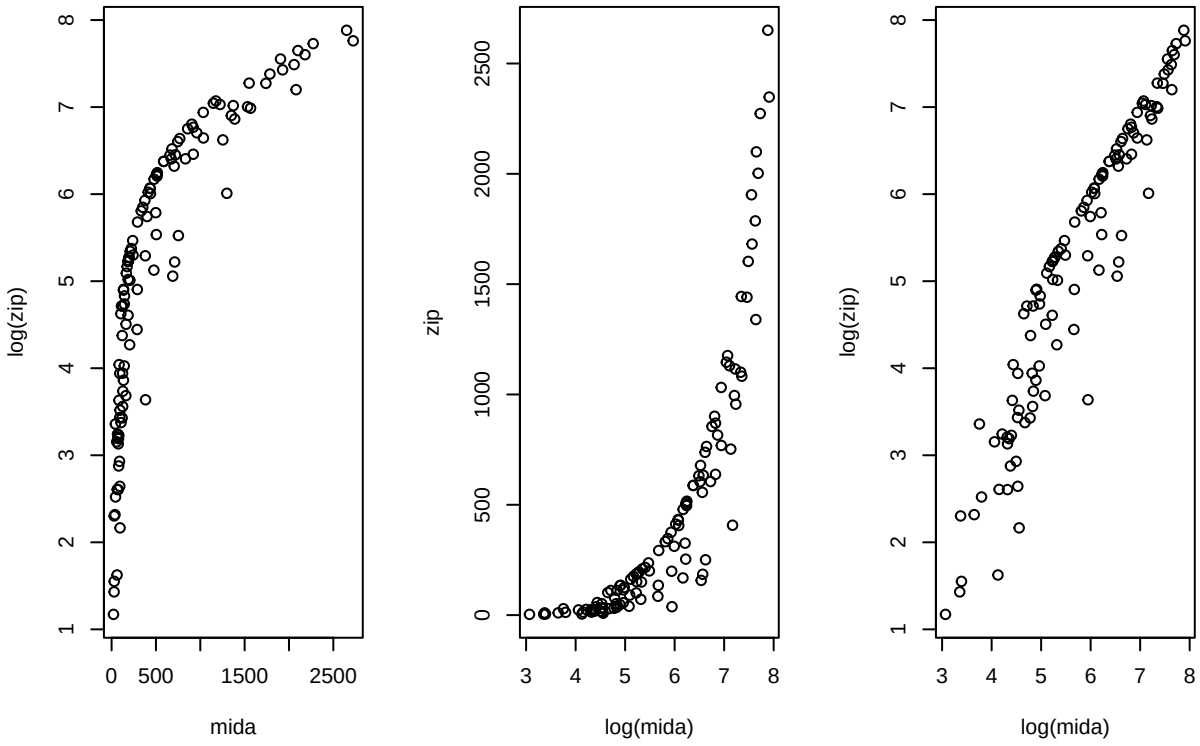
```
##
## Call:
## lm(formula = zip ~ mida, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -702.22  -28.25   18.17   75.10  345.51
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40.87691   17.93593  -2.279   0.0245 *
## mida          0.88527    0.02085  42.460   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 144.9 on 118 degrees of freedom
## Multiple R-squared:  0.9386, Adjusted R-squared:  0.938
## F-statistic:  1803 on 1 and 118 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mod_4,ask=FALSE)
```




```
par(mfrow=c(1,3))
plot(log(zip) ~ mida,     data = d)
plot(zip      ~ log(mida),data = d)
plot(log(zip) ~ log(mida),data = d)
```

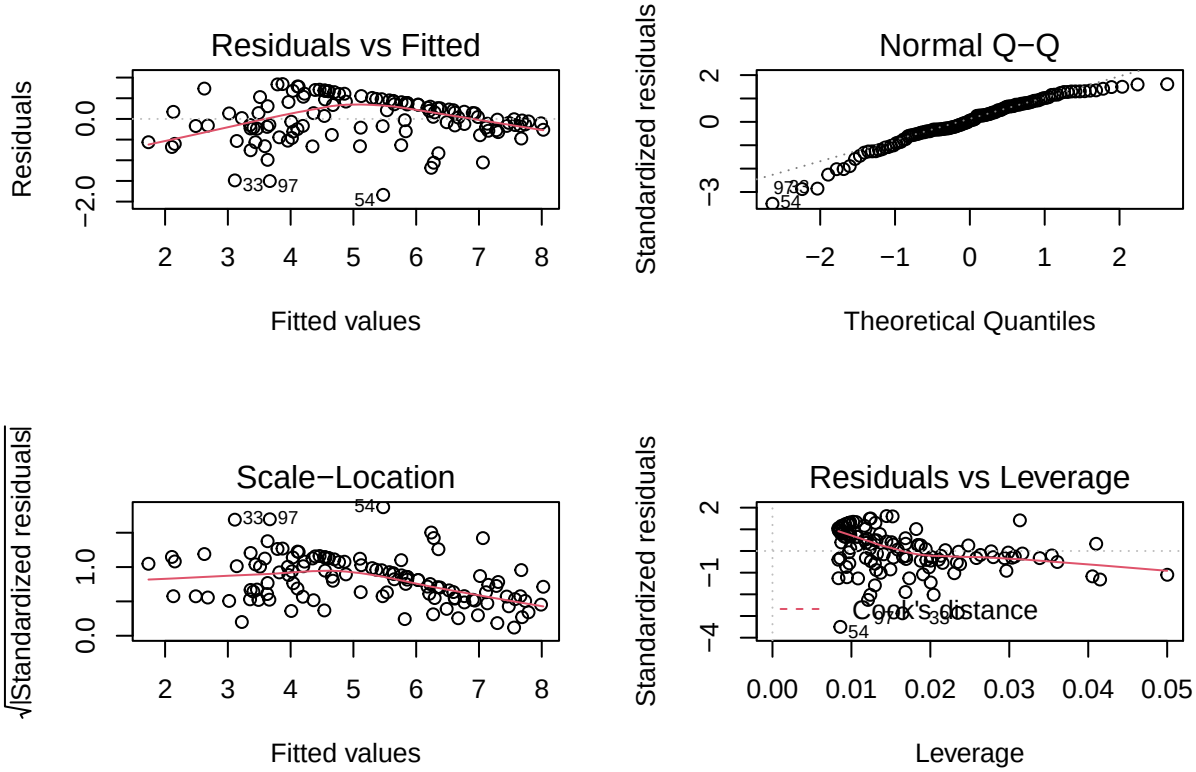



```
d$logzip  <- log(d$zip)
d$logmida <- log(d$mida)
```

```
mod_5 <- lm(logzip~logmida,data = d)
s <- summary(mod_5)
s
```

```
##
## Call:
## lm(formula = logzip ~ logmida, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83684 -0.25088  0.02739  0.39087  0.84250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.25020    0.23631  -9.522 2.72e-16 ***
## logmida      1.29896    0.04035  32.194  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5269 on 118 degrees of freedom
## Multiple R-squared:  0.8978, Adjusted R-squared:  0.8969
## F-statistic:  1036 on 1 and 118 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(mod_5,ask=FALSE)
```



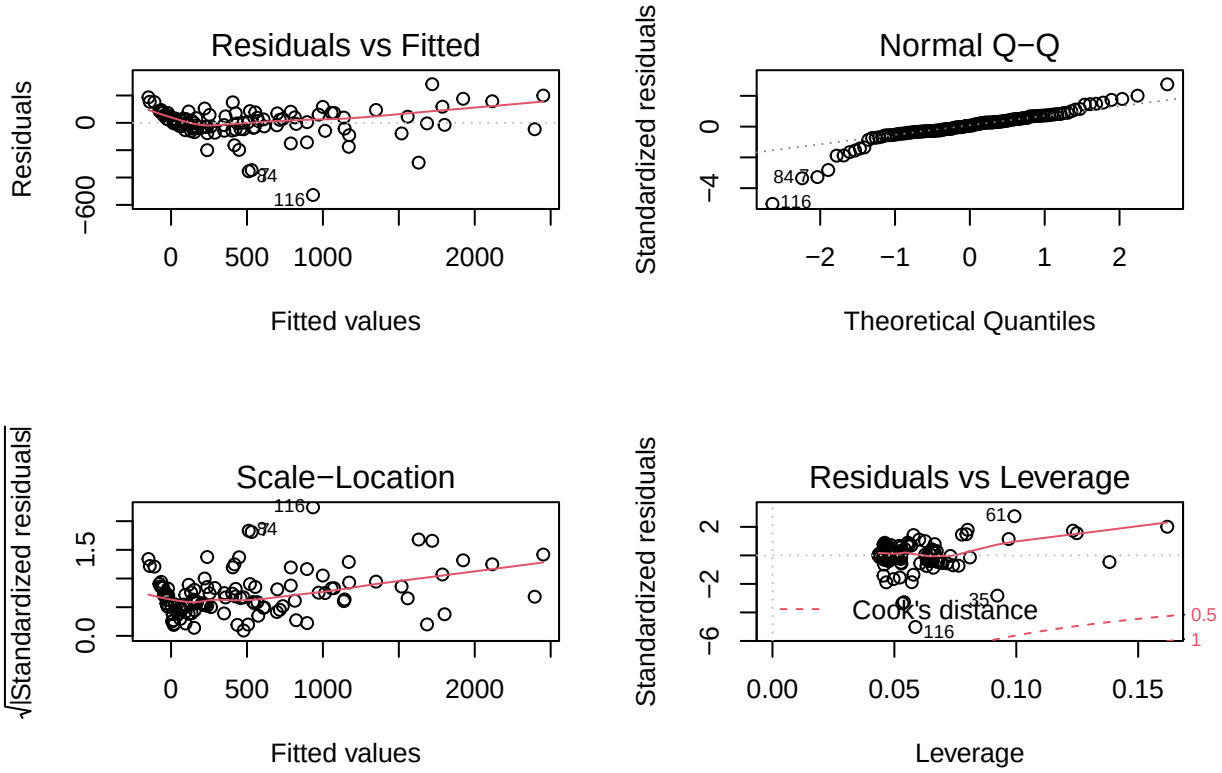

```r
##############################################################################
# Model lineal multiple (MLM)
##############################################################################
```

```r
mod_6 <- lm(zip~mida+type,data = d)
s <- summary(mod_6)
s
```

```
##
## Call:
## lm(formula = zip ~ mida + type, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -527.74  -34.57    5.96   53.48  282.51
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -49.70335   24.93911  -1.993 0.048673 *
## mida           0.89099    0.01888  47.193  < 2e-16 ***
## typejpg      100.89494   34.30685   2.941 0.003970 **
## typepdf       17.66742   37.66434   0.469 0.639919
```

```
## typepng      139.83065   39.26260   3.561 0.000541 ***
## typeppt     -173.08386   39.24493  -4.410 2.37e-05 ***
## typexls      -53.23372   33.95475  -1.568 0.119728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108.3 on 113 degrees of freedom
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9654
## F-statistic: 553.7 on 6 and 113 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mod_6,ask=FALSE)
```

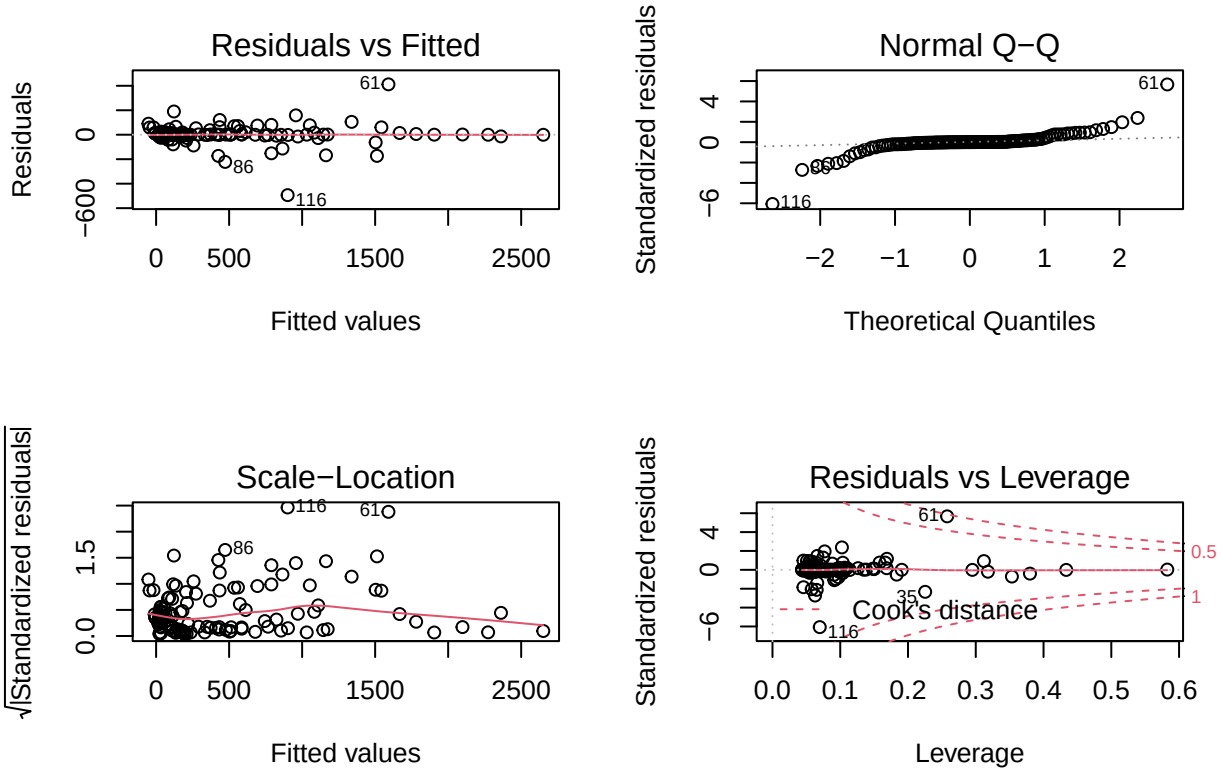



```
mod_7 <- lm(zip ~ mida*type,data = d)
s <- summary(mod_7)
s
```

```
##
## Call:
## lm(formula = zip ~ mida * type, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -493.49   -6.77    0.36   10.30  411.59
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -24.1797    40.5704  -0.596   0.5524
## mida           0.6587     0.3245   2.030   0.0448 *
## typejpg       23.6149    47.8772   0.493   0.6228
## typepdf       10.6687    50.7254   0.210   0.8338
## typepng       21.4029    51.5695   0.415   0.6789
## typeppt      -93.0354    54.7896  -1.698   0.0924 .
## typexls       30.8287    47.4251   0.650   0.5170
## mida:typejpg   0.3399     0.3268   1.040   0.3006
## mida:typepdf   0.2133     0.3255   0.655   0.5137
## mida:typepng   0.3428     0.3258   1.052   0.2951
## mida:typeppt   0.1247     0.3261   0.382   0.7030
## mida:typexls  -0.3706     0.3374  -1.099   0.2744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.29 on 108 degrees of freedom
## Multiple R-squared:  0.981,  Adjusted R-squared:  0.979
## F-statistic: 505.9 on 11 and 108 DF,  p-value: < 2.2e-16
```

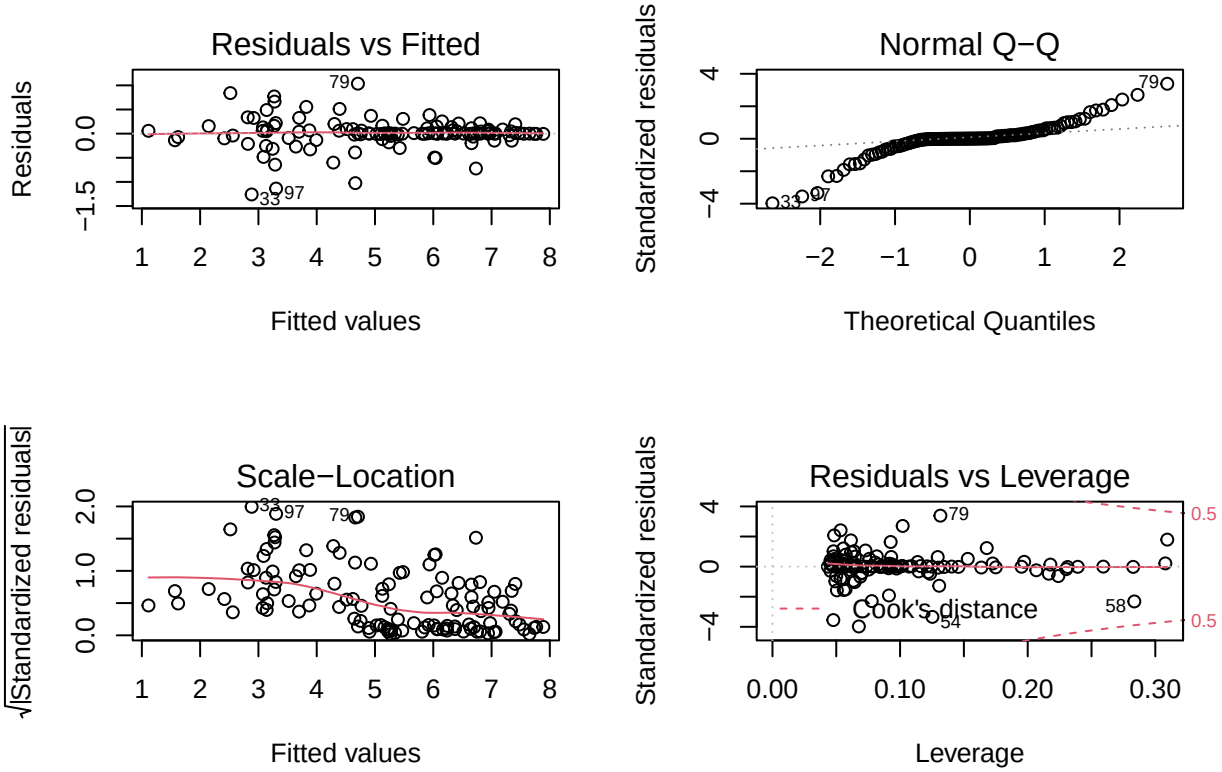
```
par(mfrow=c(2,2))
plot(mod_7,ask=FALSE)
```



```r
mod_8 <- lm(logzip ~ logmida*type,data = d)
s <- summary(mod_8)
s
```

```
##
## Call:
## lm(formula = logzip ~ logmida * type, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25976 -0.02392  0.00278  0.08411  1.03534
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.7236     0.5224  -7.127 1.21e-10 ***
## logmida           1.5765     0.1145  13.772  < 2e-16 ***
## typejpg           3.7134     0.7145   5.197 9.63e-07 ***
## typepdf           3.3906     0.7216   4.699 7.76e-06 ***
## typepng           3.6743     0.8520   4.313 3.58e-05 ***
## typeppt           1.1001     0.7337   1.499 0.136685
## typexls           2.5729     0.6442   3.994 0.000119 ***
## logmida:typejpg  -0.5753     0.1414  -4.069 9.02e-05 ***
## logmida:typepdf  -0.5525     0.1371  -4.030 0.000104 ***
## logmida:typepng  -0.5696     0.1548  -3.678 0.000368 ***
## logmida:typeppt  -0.2716     0.1384  -1.962 0.052348 .
## logmida:typexls  -0.5989     0.1384  -4.326 3.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3281 on 108 degrees of freedom
## Multiple R-squared:  0.9637, Adjusted R-squared:   0.96
## F-statistic: 260.8 on 11 and 108 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(mod_8,ask=FALSE)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage