Linear models

tommaso tarchi

2023-01-17

Contents

```
datafile <- read.csv("../WiscNursingHome.csv")
datafile <- na.omit(datafile)

data_2000 <- datafile[datafile$CRYEAR == 2000, ]
data_2001 <- datafile[datafile$CRYEAR == 2001, ]

datafile$ORGSTR <- as.factor(datafile$ORGSTR)
datafile$MSA <- as.factor(datafile$MSA)
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

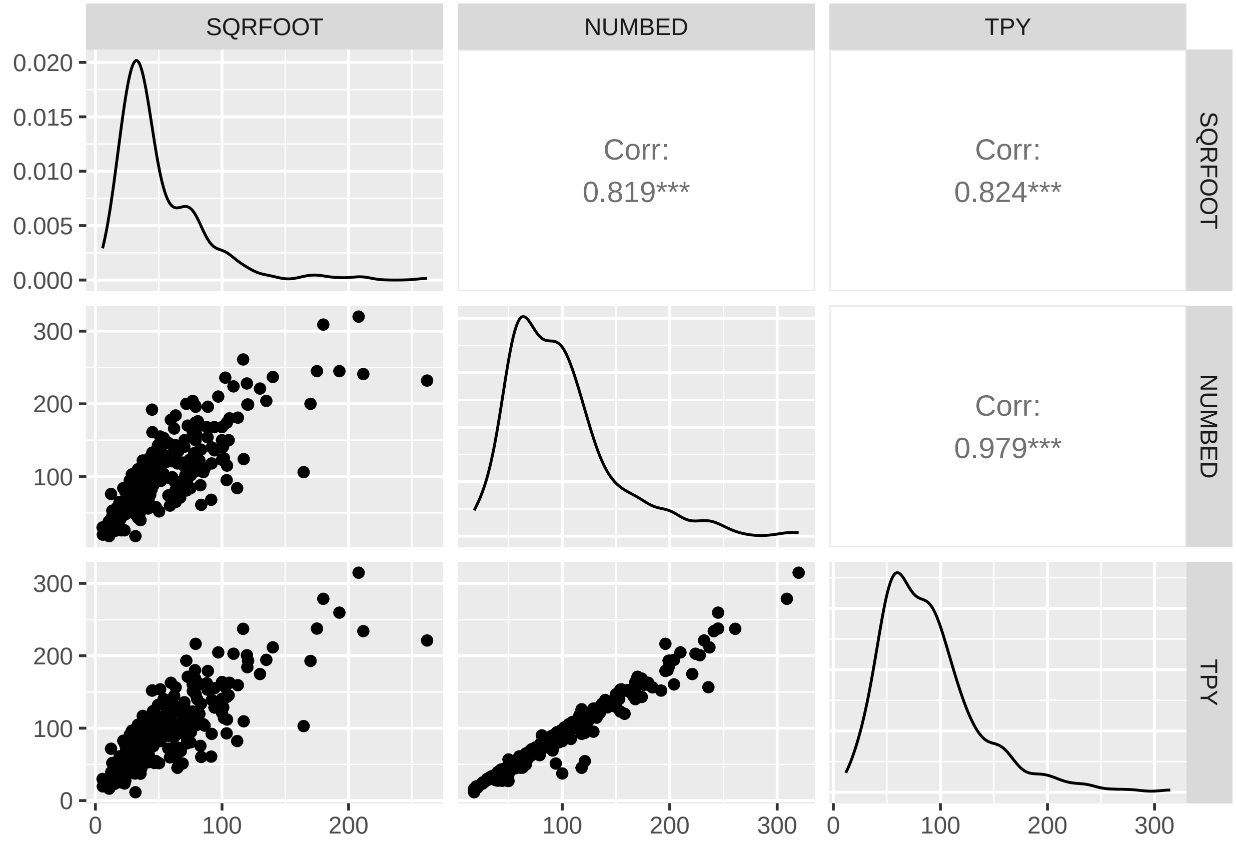
For the moment we just work on the data coming from the 2000 survey, to avoid dependency between observations.

Quantitative variables

First we consider the quantitative variables only.

Let's see the correlation:

```
ggpairs(subset(data_2000, select = c(SQRFOOT, NUMBED, TPY)))
```



We can see the strong correlation between all the three variables. In particular we see the almost perfectly linear correlation between TPY and NUMBED.

Simple linear models

Being NUMBED the most linearly correlated variable w.r.t. PTY, we start modelling PTY using NUMBED and subsequently add SQRFOOT and the interaction between the two:

```
summary(lm(TPY ~ NUMBED + SQRFOOT + NUMBED:SQRFOOT, data = data_2000))
```

```
##
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT + NUMBED:SQRFOOT, data = data_2000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.540  -1.947   0.906   4.170  39.309
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.6481214  1.7512862   0.941    0.347
## NUMBED         0.8554680  0.0202525  42.240   <2e-16 ***
## SQRFOOT        0.0528421  0.0391147   1.351    0.178
## NUMBED:SQRFOOT 0.0002358  0.0001838   1.283    0.200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 9.276 on 353 degrees of freedom
## Multiple R-squared:  0.9599, Adjusted R-squared:  0.9596
## F-statistic:  2819 on 3 and 353 DF,  p-value: < 2.2e-16
```

The t test related to SQRFOOT does not give enough evidence against the null hypotesis. To understand better the role of the variable we can perform the analysis of variance:

```
anova(glm(TPY ~ NUMBED + SQRFOOT + NUMBED:SQRFOOT, data = data_2000, family = gaussian),
      test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: TPY
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                              356     758006
## NUMBED           1   726321       355      31685 < 2.2e-16 ***
## SQRFOOT          1     1170       354      30516 0.0002268 ***
## NUMBED:SQRFOOT   1      142       353      30374 0.1995217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, despite being NUMBED and SQRFOOT highly correlated, it is worth including the second variable to the model too. Adding the interaction, instead, seems not to give any important information to the model.

We can also use Akaike information criterium to check this result (here we compare some of the possible combinations):

```
AIC(lm(TPY ~ NUMBED, data = data_2000),
    lm(TPY ~ SQRFOOT, data = data_2000),
    lm(TPY ~ NUMBED + SQRFOOT, data = data_2000),
    lm(TPY ~ NUMBED:SQRFOOT, data = data_2000),
    lm(TPY ~ NUMBED + SQRFOOT + NUMBED:SQRFOOT, data = data_2000),
    lm(TPY ~ NUMBED + NUMBED:SQRFOOT, data = data_2000),
    lm(TPY ~ SQRFOOT + NUMBED:SQRFOOT, data = data_2000))
```

```
##                                                               df      AIC
## lm(TPY ~ NUMBED, data = data_2000)                             3 2620.579
## lm(TPY ~ SQRFOOT, data = data_2000)                            3 3347.588
## lm(TPY ~ NUMBED + SQRFOOT, data = data_2000)                   4 2609.150
## lm(TPY ~ NUMBED:SQRFOOT, data = data_2000)                     3 3250.874
## lm(TPY ~ NUMBED + SQRFOOT + NUMBED:SQRFOOT, data = data_2000)  5 2609.489
## lm(TPY ~ NUMBED + NUMBED:SQRFOOT, data = data_2000)            4 2609.330
## lm(TPY ~ SQRFOOT + NUMBED:SQRFOOT, data = data_2000)           4 3250.375
```

Again the analysis indicates NUMBED as the most relevant variable and the interaction between NUMBED and SQRFOOT as basically non relevant.

Models with log-transformed predictor

We can try to improve the model by log-transforming the predictors:

```r
AIC(lm(TPY ~ log(NUMBED), data = data_2000),
    lm(TPY ~ log(SQRFOOT), data = data_2000),
    lm(TPY ~ log(NUMBED) + log(SQRFOOT), data = data_2000),
    lm(TPY ~ log(NUMBED) + SQRFOOT, data = data_2000),
    lm(TPY ~ NUMBED + log(SQRFOOT), data = data_2000))
```

```
##                                                        df      AIC
## lm(TPY ~ log(NUMBED), data = data_2000)                 3 3075.590
## lm(TPY ~ log(SQRFOOT), data = data_2000)                3 3379.167
## lm(TPY ~ log(NUMBED) + log(SQRFOOT), data = data_2000)  4 3067.573
## lm(TPY ~ log(NUMBED) + SQRFOOT, data = data_2000)       4 2956.124
## lm(TPY ~ NUMBED + log(SQRFOOT), data = data_2000)       4 2618.490
```

We cannot see any improvement.

Graphical analysis

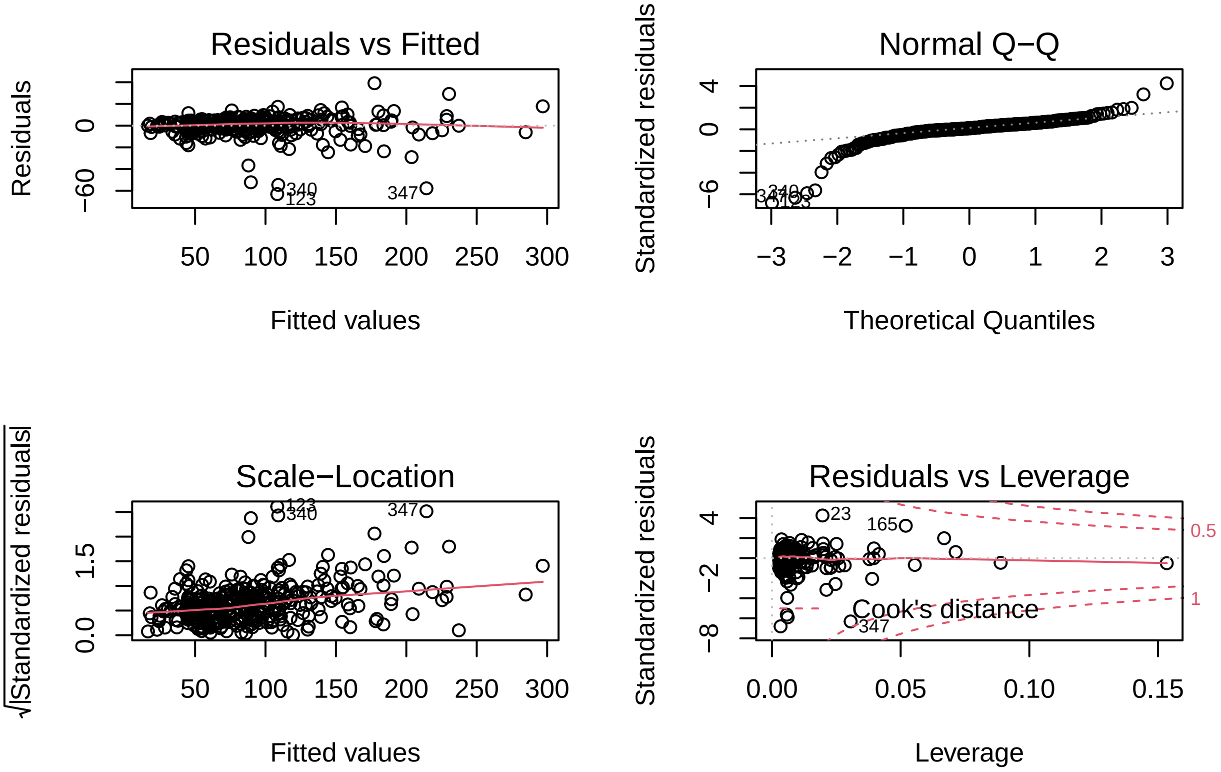
At the moment, the best model seems to be the one with just the single linear contributions of the two variables. Let's analyze the residuals:

```r
fit.linear <- lm(TPY ~ NUMBED + SQRFOOT, data = data_2000)

summary(fit.linear)
```

```
## 
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT, data = data_2000)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -63.052  -1.764   0.952   4.256  39.073 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.09458    1.10635  -0.085 0.931922    
## NUMBED       0.86858    0.01750  49.627  < 2e-16 ***
## SQRFOOT      0.09160    0.02487   3.684 0.000266 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.285 on 354 degrees of freedom
## Multiple R-squared:  0.9597, Adjusted R-squared:  0.9595 
## F-statistic:  4220 on 2 and 354 DF,  p-value: < 2.2e-16
```

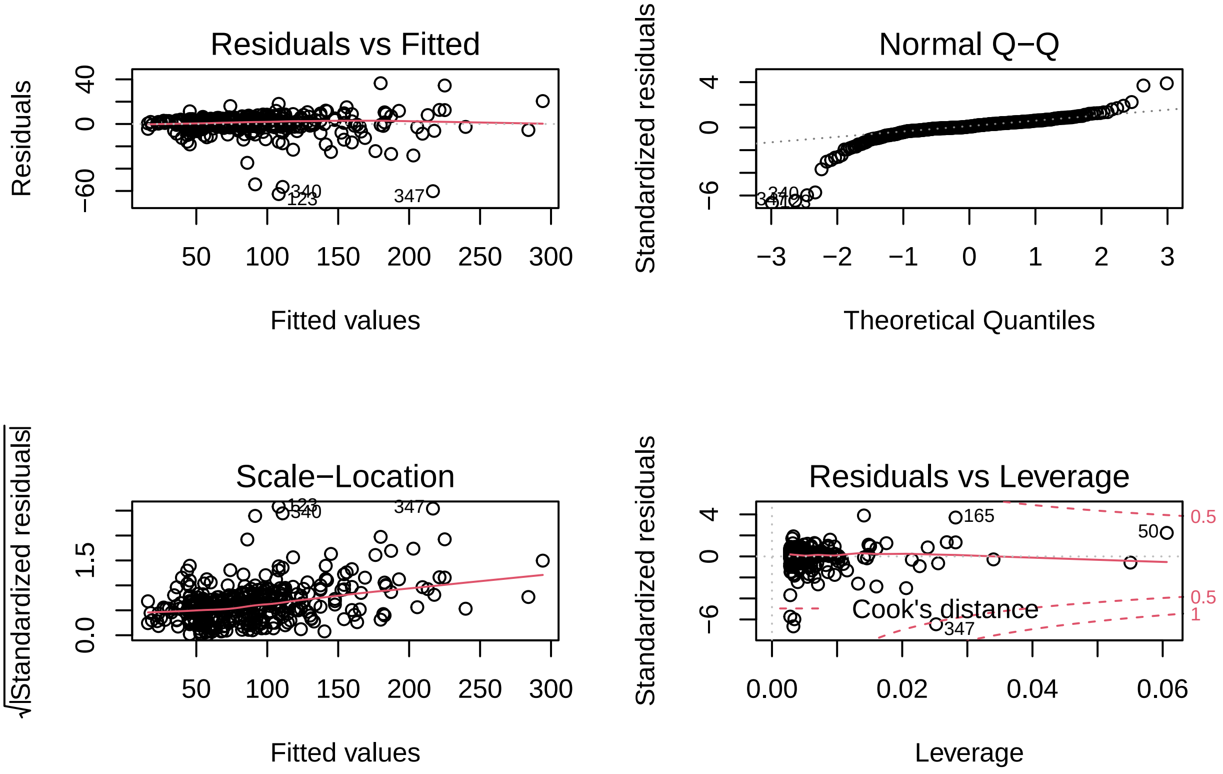
```r
par(mfrow = c(2, 2))
plot(fit.linear)
```



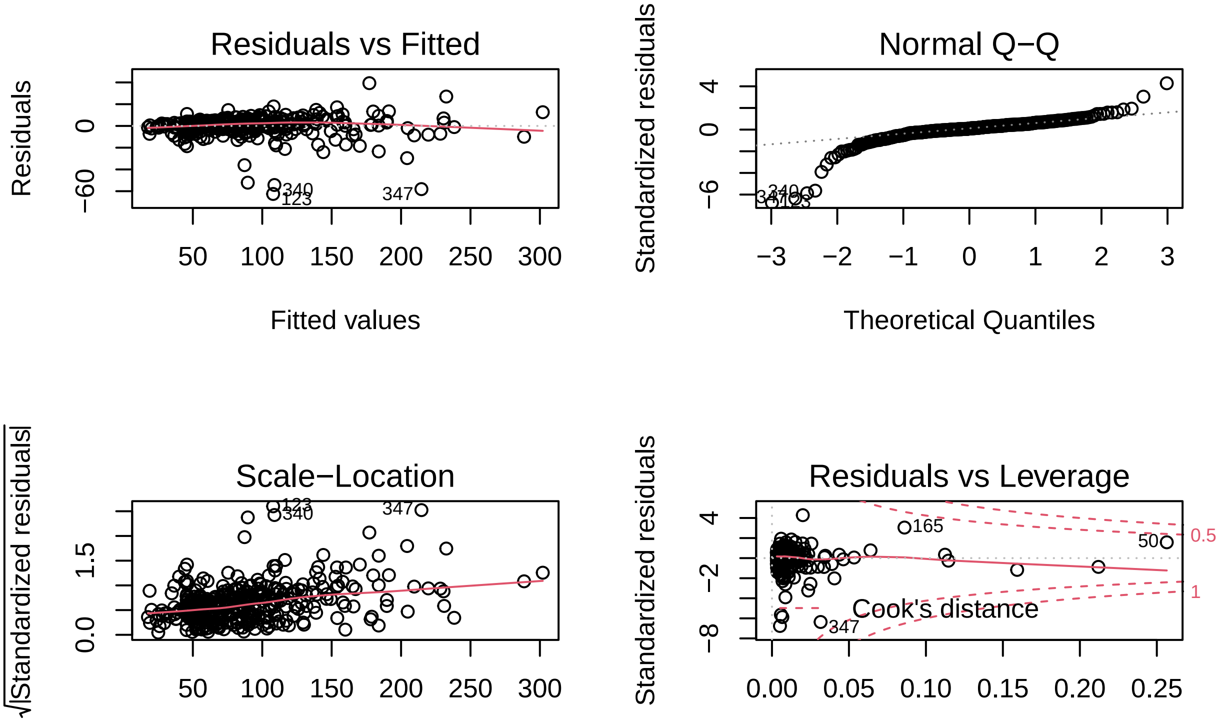
The residuals are not exactly as we would expect from a good linear model. In particular there seem to be a couple of outliers, the residuals are not normally distributed on the edges and homoscedasticity is not satisfied.

Let's also try to inspect the residuals' plots for the other good models (according to AIC) we got previously:

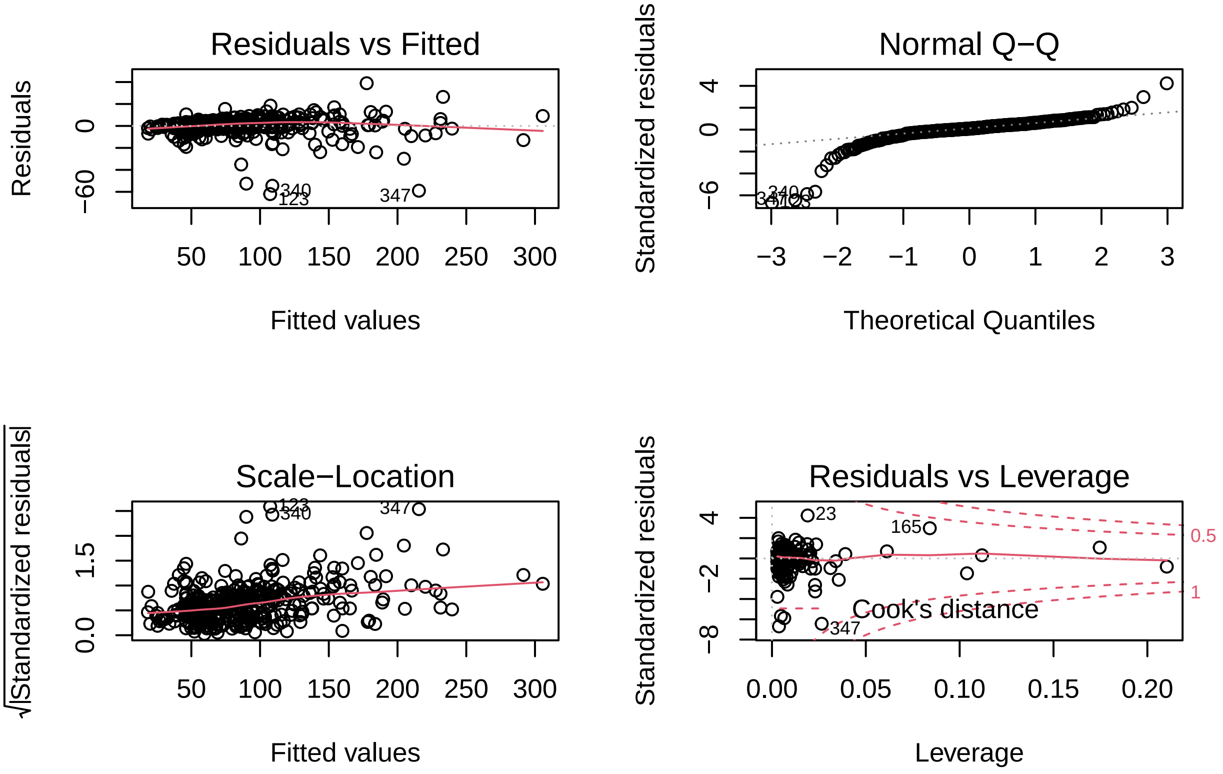
```
par(mfrow = c(2, 2))
plot(lm(TPY ~ NUMBED, data = data_2000))
```



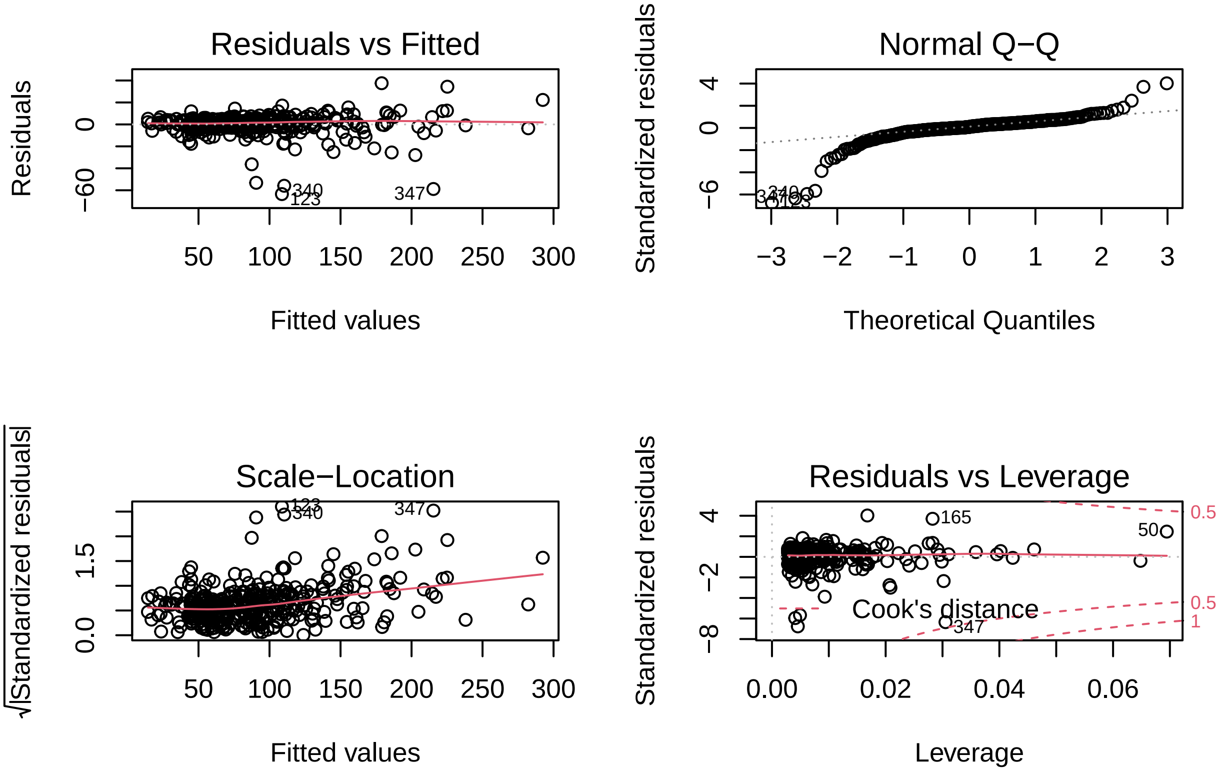
```
par(mfrow = c(2, 2))
plot(lm(TPY ~ NUMBED + SQRFOOT + NUMBED:SQRFOOT, data = data_2000))
```



```
par(mfrow = c(2, 2))
plot(lm(TPY ~ NUMBED + NUMBED:SQRFOOT, data = data_2000))
```



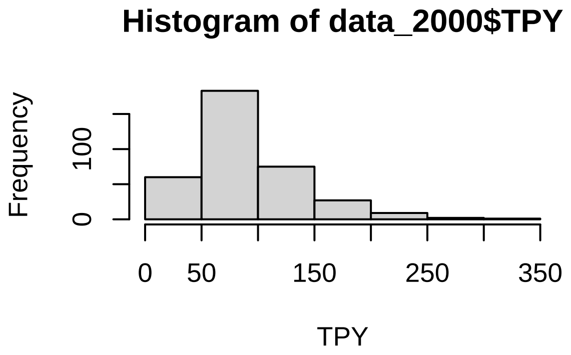
```r
par(mfrow = c(2, 2))
plot(lm(TPY ~ NUMBED + log(SQRFOOT), data = data_2000))
```

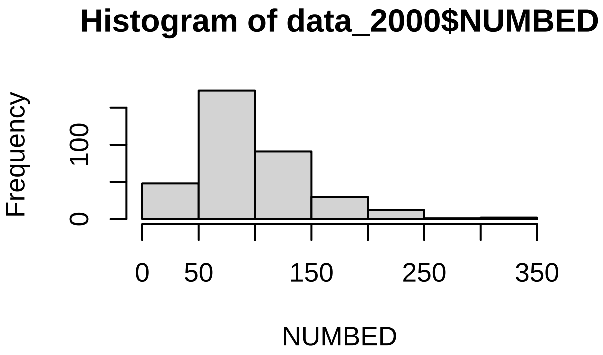


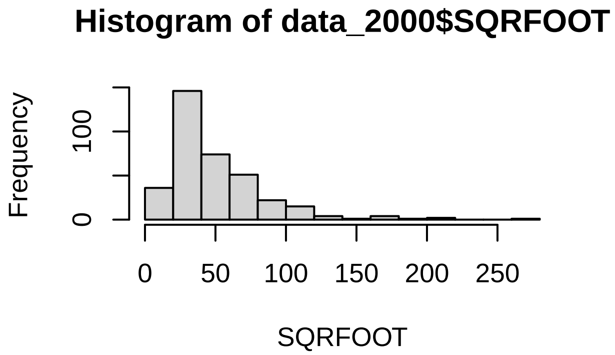
None of the models above seems to satisfy the assumptions on the linear model.

Let's now look at the distribution of the data:

```
par(mfrow = c(2, 2))
hist(data_2000$TPY, xlab = "TPY", freq = TRUE)
hist(data_2000$NUMBED, xlab = "NUMBED", freq = TRUE)
hist(data_2000$SQRFOOT, xlab = "SQRFOOT", freq = TRUE)
```







As we can see all of the variables are strongly skewed.

Models with log-transformed TPY

We see that also TPY is strongly skewed, therefore we can try to model its log-transform:

```
summary(lm(log(TPY) ~ log(NUMBED) + log(SQRFOOT) + log(NUMBED):log(SQRFOOT), data = data_2000))
```

```
##
## Call:
## lm(formula = log(TPY) ~ log(NUMBED) + log(SQRFOOT) + log(NUMBED):log(SQRFOOT),
##     data = data_2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87621 -0.01560  0.01771  0.05989  0.23298
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.071165   0.251030  -0.283    0.777
## log(NUMBED)               0.973814   0.059765  16.294   <2e-16 ***
## log(SQRFOOT)              0.008461   0.069971   0.121    0.904
## log(NUMBED):log(SQRFOOT)  0.003516   0.015026   0.234    0.815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1169 on 353 degrees of freedom
## Multiple R-squared:  0.9478, Adjusted R-squared:  0.9474
## F-statistic:  2139 on 3 and 353 DF,  p-value: < 2.2e-16
```

```
anova(glm(log(TPY) ~ log(NUMBED) + log(SQRFOOT) + log(NUMBED):log(SQRFOOT), data = data_2000,
          family = gaussian), test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: log(TPY)
##
## Terms added sequentially (first to last)
##
##
##                           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                        356     92.570
## log(NUMBED)                1   87.718       355      4.852   <2e-16 ***
## log(SQRFOOT)               1    0.023       354      4.828   0.1902
## log(NUMBED):log(SQRFOOT)   1    0.001       353      4.828   0.8150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case even the main effect of SQRFOOT seems to be non relevant. Let's give a look at the plot of the model with just log(NUMBED) and with just log(SQRFOOT):

```
summary(lm(log(TPY) ~ log(NUMBED), data = data_2000))
```

```
##
## Call:
## lm(formula = log(TPY) ~ log(NUMBED), data = data_2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88769 -0.01480  0.01781  0.06355  0.23079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.15044    0.05672  -2.652  0.00835 **
## log(NUMBED)  1.01194    0.01263  80.114  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1169 on 355 degrees of freedom
## Multiple R-squared:  0.9476, Adjusted R-squared:  0.9474
## F-statistic:  6418 on 1 and 355 DF,  p-value: < 2.2e-16
```

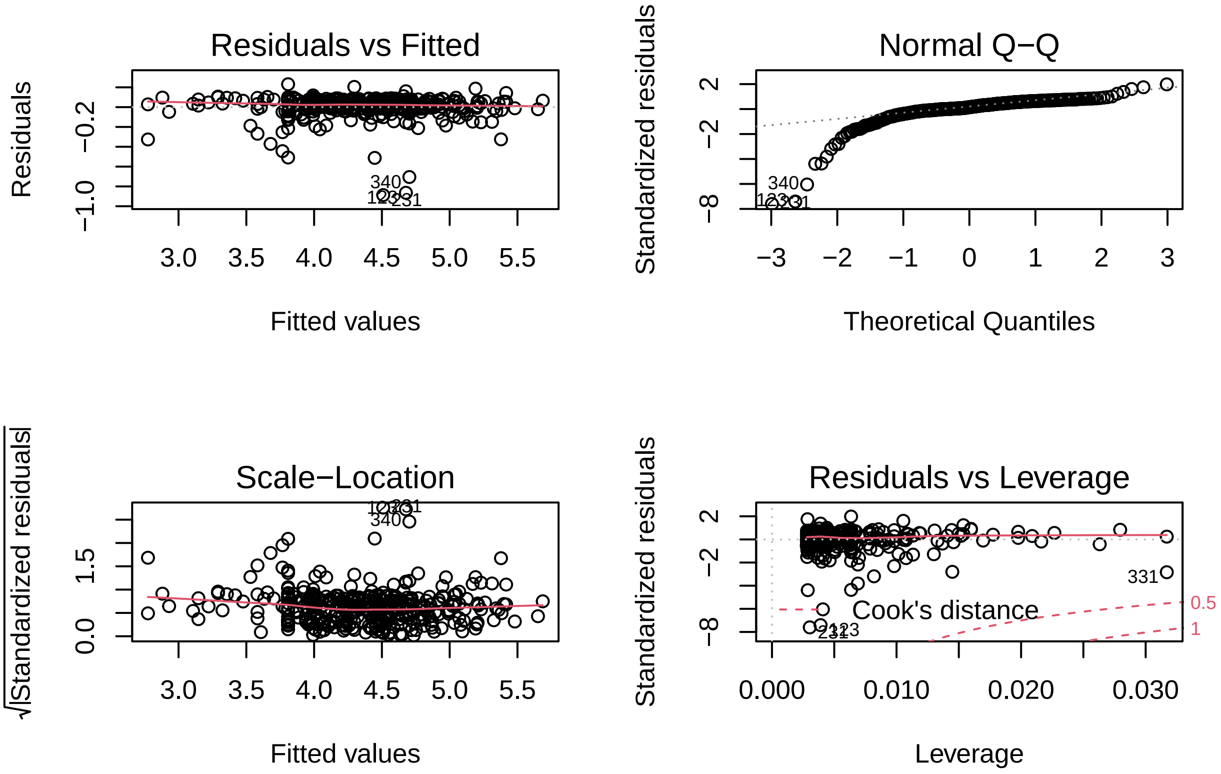
```
summary(lm(log(TPY) ~ log(SQRFOOT), data = data_2000))
```

```
##
## Call:
## lm(formula = log(TPY) ~ log(SQRFOOT), data = data_2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73074 -0.14273  0.02319  0.19852  0.73846
##
```

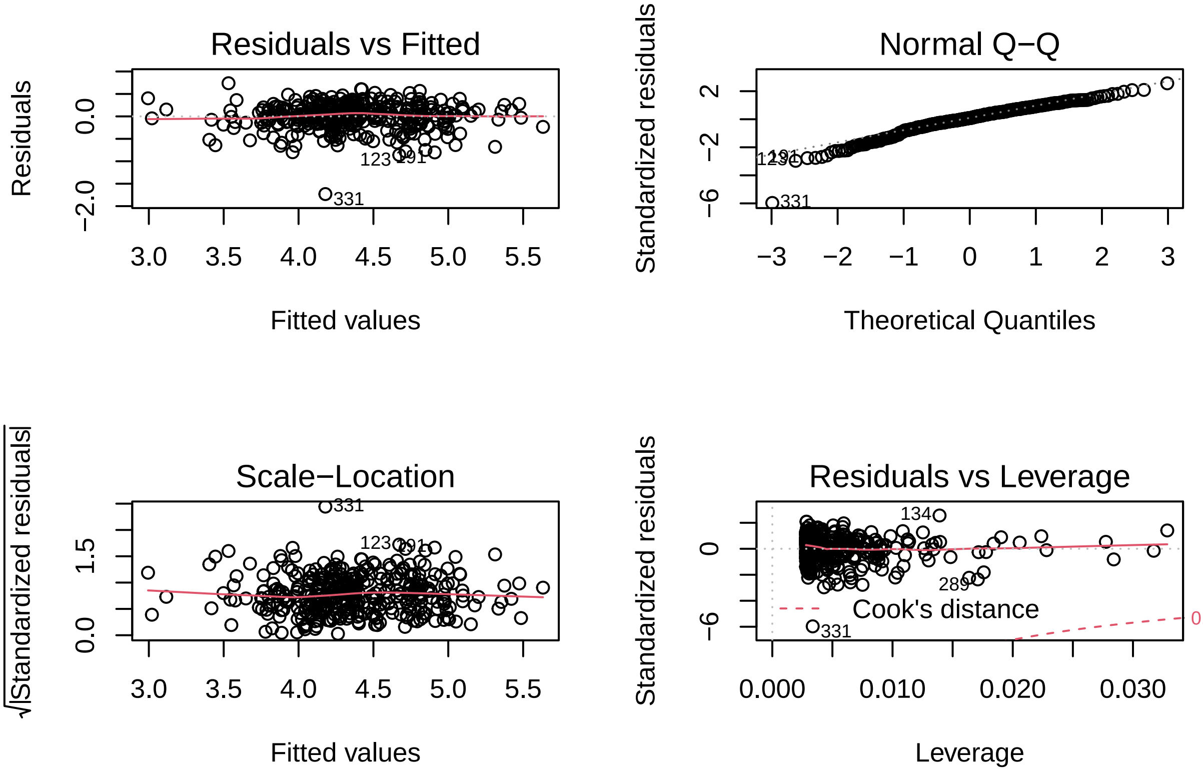
```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     1.80509    0.09525   18.95   <2e-16 ***
## log(SQRFOOT)    0.68737    0.02523   27.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2904 on 355 degrees of freedom
## Multiple R-squared:  0.6765, Adjusted R-squared:  0.6756 
## F-statistic: 742.5 on 1 and 355 DF,  p-value: < 2.2e-16
```

And to their residuals:

```r
par(mfrow = c(2, 2))
plot(lm(log(TPY) ~ log(NUMBED), data = data_2000))
```



```r
par(mfrow = c(2, 2))
plot(lm(log(TPY) ~ log(SQRFOOT), data = data_2000))
```



```
AIC(lm(log(TPY) ~ log(NUMBED), data = data_2000), lm(log(TPY) ~ log(SQRFOOT), data = data_2000))
```

```
##                                                df       AIC
## lm(log(TPY) ~ log(NUMBED), data = data_2000)    3 -515.4005
## lm(log(TPY) ~ log(SQRFOOT), data = data_2000)   3  134.3160
```

Again NUMBED seems to be a good predictor, but the residuals suggest that the assumption on the linear model are not satisfied.

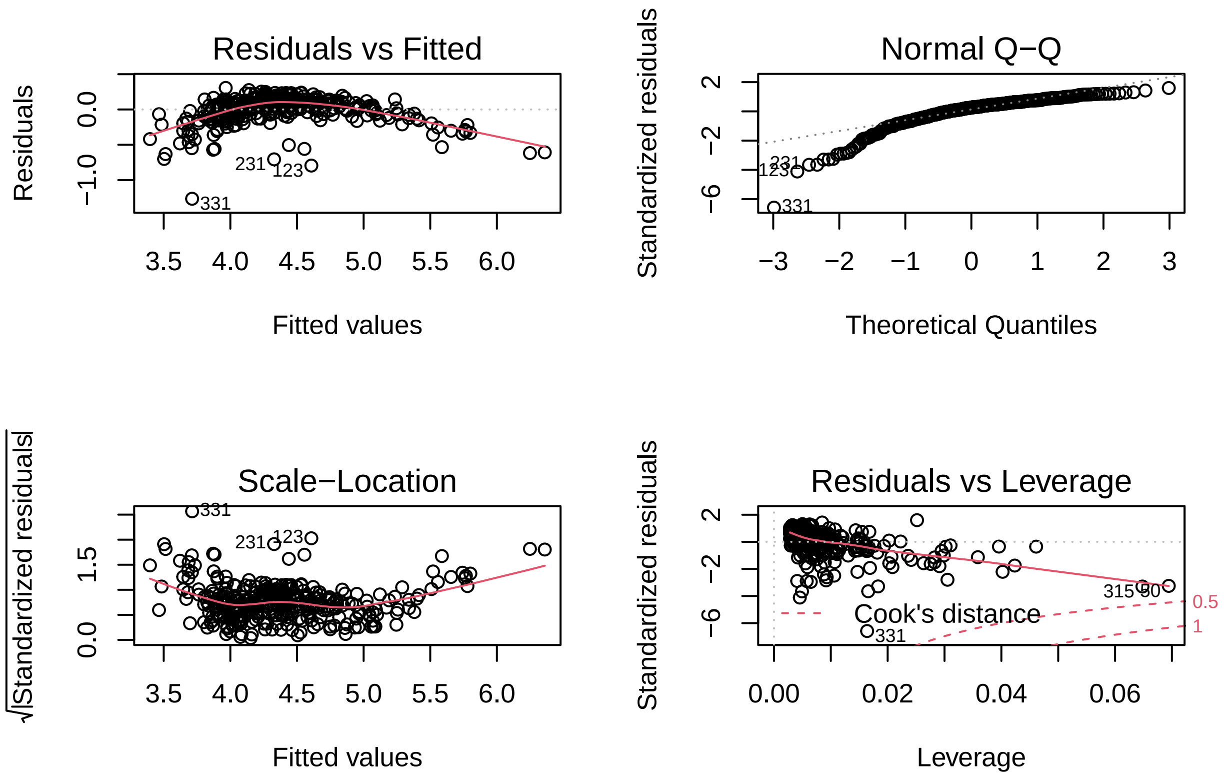
On the other hand SQRFOOT seems to be less relevant but with mush better residuals (apart from the high leverage outlier 331).

We can also try a kind of mixed model:

```
summary(lm(log(TPY) ~ NUMBED + log(SQRFOOT), data = data_2000))
```

```
##
## Call:
## lm(formula = log(TPY) ~ NUMBED + log(SQRFOOT), data = data_2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26488 -0.07008  0.05039  0.12170  0.30548
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.8984517  0.0820220  35.337  < 2e-16 ***
## NUMBED       0.0075357  0.0003576  21.073  < 2e-16 ***
## log(SQRFOOT) 0.1966199  0.0287299   6.844 3.41e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1937 on 354 degrees of freedom
## Multiple R-squared:  0.8565, Adjusted R-squared:  0.8557
## F-statistic:  1057 on 2 and 354 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(lm(log(TPY) ~ NUMBED + log(SQRFOOT), data = data_2000))
```



```
AIC(lm(log(TPY) ~ NUMBED + log(SQRFOOT), data = data_2000))
```

```
## [1] -153.8819
```

But the results are not encouraging.