Linear models

2023-04-29

Exercises

Using the data set Soils:

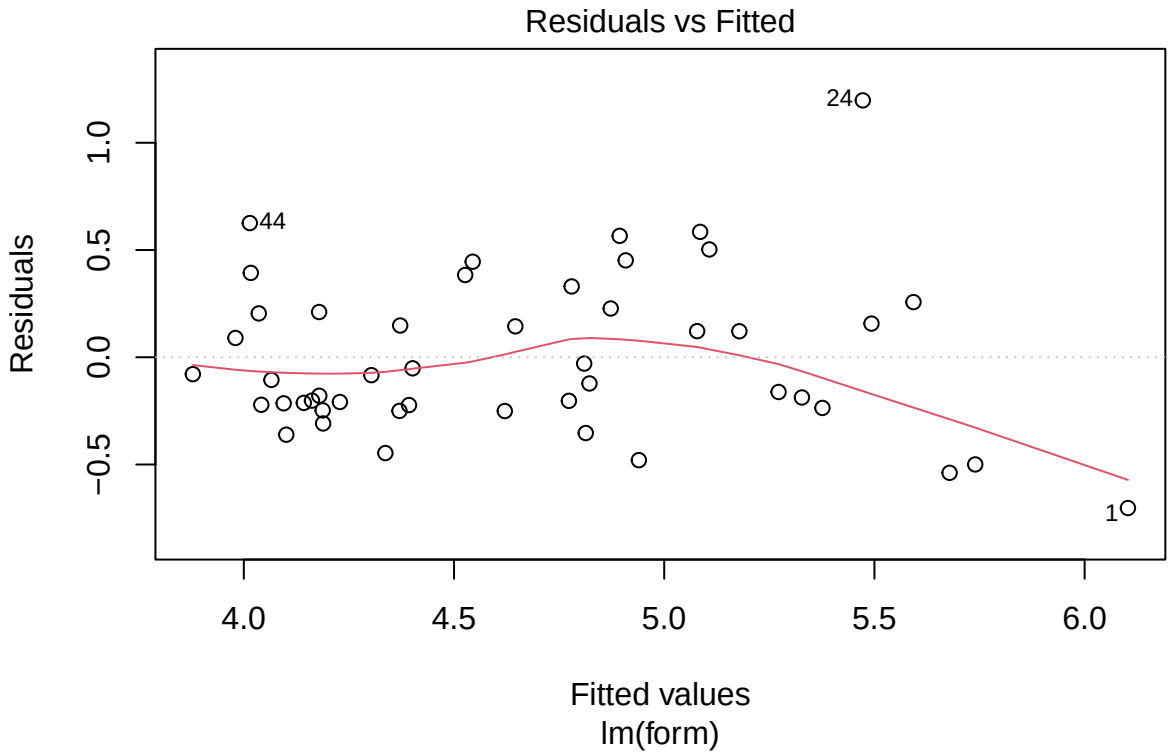
1 Fit the following model: $\text{pH} = 0 + 1\,\text{N} + 2\,\text{P} + 3\,\text{Ca} + 4\,\text{Mg} + 5\,\text{K} + 6\,\text{Na} +$

```
# Predictor variables
pred_vars <- c("N", "P", "Ca", "Mg", "K", "Na")

# Create the equation
form <- str_c("pH", " ~ ", str_c(pred_vars, collapse = " + ")) %>%
  as.formula()

# Fit the model
m1 <- lm(formula = form, data = Soils)

summary(m1)
```

```
##
## Call:
## lm(formula = form, data = Soils)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70302 -0.22629 -0.09459  0.21499  1.19768
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.801134   0.597108   6.366 1.31e-07 ***
## N           -5.028408   2.261120  -2.224   0.0317 *
## P            0.001793   0.001407   1.275   0.2095
## Ca           0.193754   0.037999   5.099 8.15e-06 ***
## Mg          -0.029468   0.051959  -0.567   0.5737
## K           -0.056504   0.390086  -0.145   0.8855
## Na          -0.035206   0.036402  -0.967   0.3391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3942 on 41 degrees of freedom
## Multiple R-squared:  0.6996, Adjusted R-squared:  0.6557
## F-statistic: 15.92 on 6 and 41 DF,  p-value: 2.412e-09
```

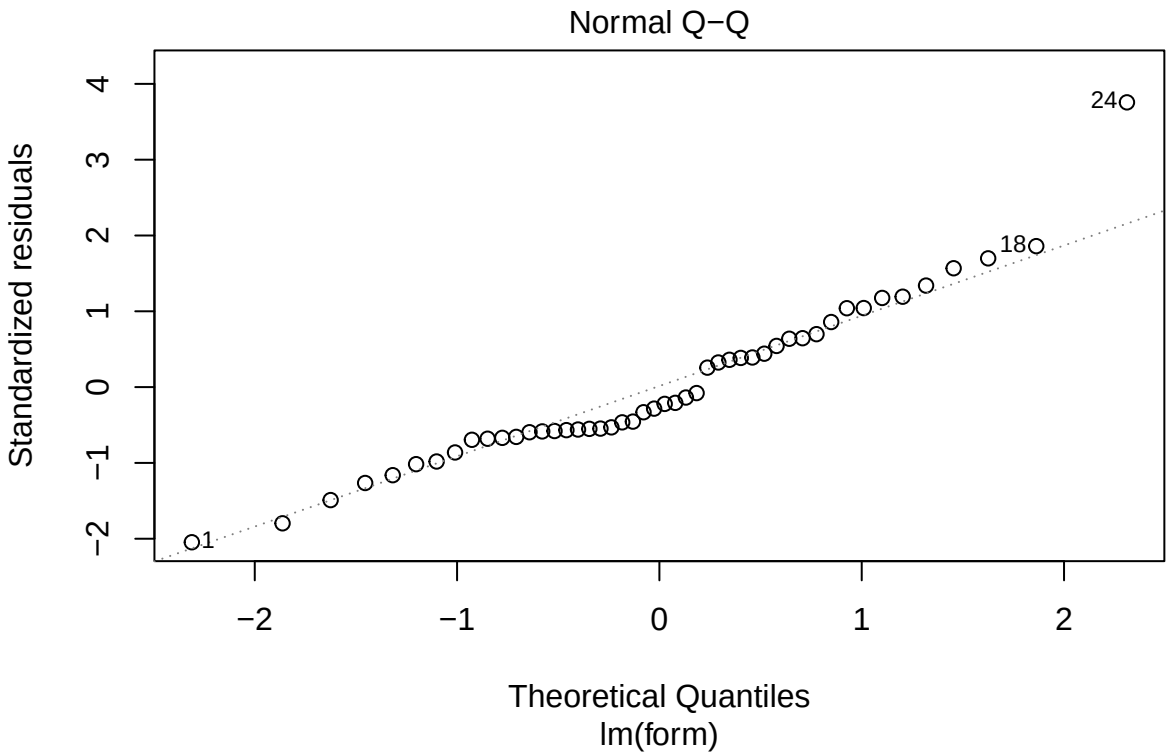
2 Make the analysis of diagnostics - plot the residuals and identify possible leverage points and outliers. Below we have the residuals, or error. In the y axis is expected that the values fluctuate around the 0, since we want to minimize the error. In the x axis between 4 and 5 we have values with low error, however, after 5, we have the tendency line being down skewed. It might be an indication of heterocedasticity.

```
plot(m1, which = 1)
```



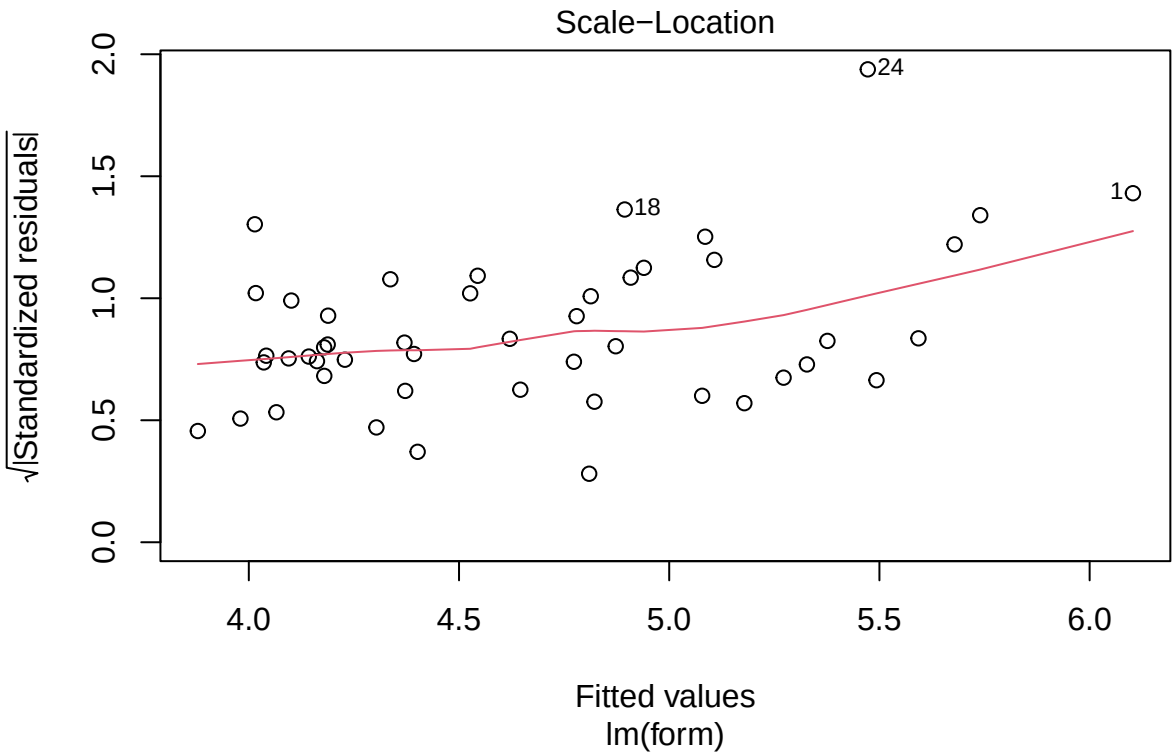
Below we have a qq-plot that helps to identifiy if our residuals are normal distributed. Errors are expecte to be random and consequently follows a normal distribution. Below our residuals tend to follow the dashed line, which suggests that they might be normal distributed.

```
plot(m1, which = 2)
```



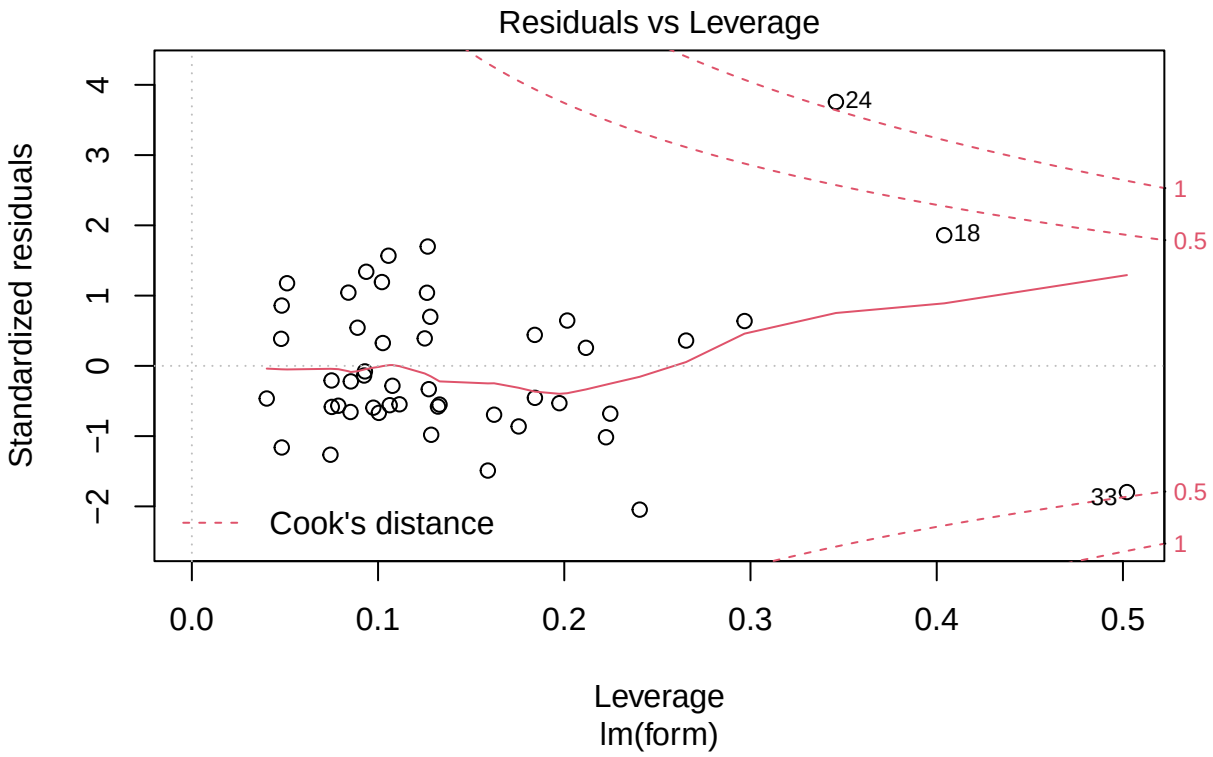
Standardized residuals are difficult to interpret, it's easy to understand based on not standardized.

```
plot(m1, which = 3)
```



Cook's distance are useful to detect outliers and leverage points. Below we see the samples 18, 24 and 33 highlighted. Dot 18 and 33 is still inside where the biggest amount of the data are, than it's ok. Dot 24, however is above 1 standardized residuals, suggesting that this point influence the model. Then we could remove it and fit another model to check its influence.

```
plot(m1, which = 5)
```



3 Based on ANOVA, is there any variable that is not significant to explain pH? Our anova results suggests that P, Mg, K and Na are not statistically important to explain pH.

```
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: pH
##           Df Sum Sq Mean Sq F value    Pr(>F)
## N          1 8.5992  8.5992 55.3279 3.993e-09 ***
## P          1 0.2195  0.2195  1.4122    0.2415
## Ca         1 5.7509  5.7509 37.0022 3.297e-07 ***
## Mg         1 0.1260  0.1260  0.8108    0.3732
## K          1 0.0020  0.0020  0.0131    0.9096
## Na         1 0.1454  0.1454  0.9354    0.3391
## Residuals 41 6.3723  0.1554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 What happens when we fit the following model: $\text{pH} = 1\,\text{N} + 2\,\text{P} + 3\,\text{Ca} + 4\,\text{Mg} + 5\,\text{K} + 6\,\text{Na} + ?$

Repeat analyses 1, 2 and 3.

Briefly, we observe variables that were important in m1, are not statistically important anymore to expain pH when compared to m2, except for Ca. This was in consequence of intercept removal.

```r
# Predictor variables
pred_vars <- c(-1, "N", "P", "Ca", "Mg", "K", "Na")

# Create the equation
form <- str_c("pH", " ~ ", str_c(pred_vars, collapse = " + ")) %>%
  as.formula()

# Fit the model
m2 <- lm(formula = form, data = Soils)

summary(m2)
```

```
##
## Call:
## lm(formula = form, data = Soils)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0507 -0.3429  0.1226  0.3609  0.9855
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## N  -5.218591   3.149966  -1.657  0.10503
## P   0.005319   0.001801   2.953  0.00514 **
## Ca  0.266696   0.050476   5.284 4.22e-06 ***
## Mg  0.174643   0.056964   3.066  0.00379 **
## K   0.457001   0.531730   0.859  0.39497
## Na  0.080400   0.043954   1.829  0.07448 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5493 on 42 degrees of freedom
## Multiple R-squared:  0.9881, Adjusted R-squared:  0.9864
## F-statistic: 582.9 on 6 and 42 DF,  p-value: < 2.2e-16
```

5 Now include Block effects. What can we see? With the addition of Block to our model, Block 2 and 4 are statistically important to explain the variability in soils data, besides intercept, N, Ca.

```r
# Predictor variables
pred_vars <- c("N", "P", "Ca", "Mg", "K", "Na", "Block")

# Create the equation
form <- str_c("pH", " ~ ", str_c(pred_vars, collapse = " + ")) %>%
  as.formula()

# Fit the model
m3 <- lm(formula = form, data = Soils)

summary(m3)
```

```
##
## Call:
```

```
## lm(formula = form, data = Soils)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67903 -0.15402  0.01854  0.13022  0.99782
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.177373   0.694000   4.578 4.91e-05 ***
## N           -5.235253   1.942666  -2.695 0.010428 *
## P            0.002142   0.001240   1.727 0.092324 .
## Ca           0.208774   0.034456   6.059 4.72e-07 ***
## Mg           0.030221   0.048852   0.619 0.539850
## K           -0.529399   0.403474  -1.312 0.197357
## Na          -0.064927   0.034604  -1.876 0.068308 .
## Block2       0.596972   0.156152   3.823 0.000476 ***
## Block3       0.254605   0.169516   1.502 0.141374
## Block4       0.540492   0.139029   3.888 0.000394 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3284 on 38 degrees of freedom
## Multiple R-squared:  0.8069, Adjusted R-squared:  0.7611
## F-statistic: 17.64 on 9 and 38 DF,  p-value: 5.03e-11
```

6 Is there any significant block effect? Yes, blocks 2 and 4

7 Is there any variable deviating from normality? Is there any transformation possible? For all scenarios below, I would assume that all variables don't deviate from normality, based on the shapiro test.

```
bind_rows(
  shapiro.test(residuals(m1)) %>% broom::tidy() %>% mutate(model = "m1"),
  shapiro.test(residuals(m2)) %>% broom::tidy() %>% mutate(model = "m2"),
  shapiro.test(residuals(m3)) %>% broom::tidy() %>% mutate(model = "m3")
)
```

```
## # A tibble: 3 x 4
##   statistic p.value method                    model
##       <dbl>   <dbl> <chr>                     <chr>
## 1     0.952  0.0464 Shapiro-Wilk normality test m1
## 2     0.966  0.181  Shapiro-Wilk normality test m2
## 3     0.958  0.0823 Shapiro-Wilk normality test m3
```

8 Is there some normalization that we could apply? Since all models are normal distributed, we don't need to apply any normalization.