CODICEPERSONA_PROBLEMA

Marco Scarpelli

DATA

Print dataframe

```
##    rate rain hardness coarse  fine
## 1 31.04  647        4   9.81  8.93
## 2 31.05  689        5  12.62 10.29
## 3 30.57  715        7  10.55  8.76
## 4 30.95  661        4  12.09  8.39
## 5 32.53  677        4  12.55 10.51
## 6 29.57  660        4  10.40  7.79
```

```
## [1] 80  5
```

Point a

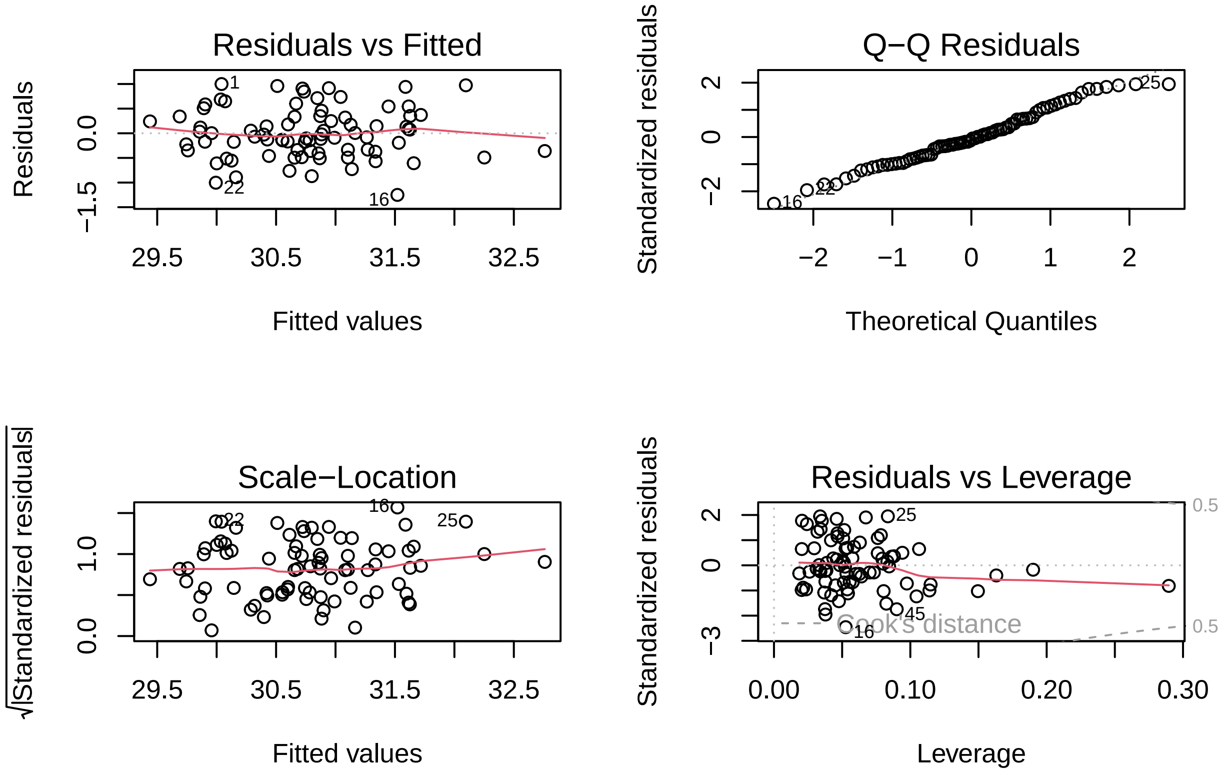
```
##
## Call:
## lm(formula = rate ~ rain + hardness + coarse + fine, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25053 -0.35960 -0.04854  0.34265  0.99875
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.514647   1.006933  20.373  < 2e-16 ***
## rain         0.006115   0.001033   5.922 8.96e-08 ***
## hardness     0.011423   0.049766   0.230    0.819
## coarse       0.385217   0.047864   8.048 9.67e-12 ***
## fine         0.195448   0.046351   4.217 6.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5225 on 75 degrees of freedom
## Multiple R-squared:  0.6183, Adjusted R-squared:  0.5979
## F-statistic: 30.37 on 4 and 75 DF,  p-value: 5.008e-15
```

```
##  (Intercept)         rain     hardness       coarse         fine
## 20.514647296  0.006115268  0.011423423  0.385216816  0.195448455
```

Our assumptions are that the residuals have 0 mean and are homoscedastic. Let us check for their Gaussianity:

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  residuals(fm)
## W = 0.98286, p-value = 0.3607
```



Furthermore, we can see that the Q-Q plot follows the line closely enough, and all points on the residual-leverage plot are within Cook's distance.

Let us also check the variance inflation factor:

```
##     rain hardness   coarse     fine
## 1.016423 1.023157 1.043175 1.048110
```

where we can see that all parameters are well below 5 and especially 10.

Point b

From the summary, we can see that the variable `hardness` is strongly not significant for our model. We will remove it:

```
##
## Call:
## lm(formula = rate ~ rain + coarse + fine, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25732 -0.37283 -0.04774  0.34202  0.99202
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 20.560257   0.980962  20.959  < 2e-16 ***
## rain         0.006132   0.001024   5.989 6.54e-08 ***
## coarse       0.386020   0.047438   8.137 6.01e-12 ***
## fine         0.194142   0.045713   4.247 6.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5193 on 76 degrees of freedom
## Multiple R-squared:  0.618,  Adjusted R-squared:  0.6029 
## F-statistic: 40.98 on 3 and 76 DF,  p-value: 7.312e-16
```

```
##  (Intercept)         rain       coarse         fine 
## 20.560257374  0.006131644  0.386020255  0.194141815
```

where we can see that now all parameters seem to be significant for our model. Let us check whether the two models are equivalent:

```
## Analysis of Variance Table
## 
## Model 1: rate ~ rain + hardness + coarse + fine
## Model 2: rate ~ rain + coarse + fine
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     75 20.478                           
## 2     76 20.492 -1 -0.014387 0.0527 0.8191
```

According to the output's p-value (0.81), the two models perform equal predictions.

Point c

```
## Linear hypothesis test
## 
## Hypothesis:
## coarse - 2 fine = 0
## 
## Model 1: restricted model
## Model 2: rate ~ rain + coarse + fine
## 
##   Res.Df    RSS Df  Sum of Sq     F Pr(>F)
## 1     77 20.492                           
## 2     76 20.492  1 0.00011469 4e-04 0.9836
```

The null hypothesis is thus proven; we can update the model dataframe with a new column to account for this.

Attempt 1

> Dubbio: qui secondo me avrebbe senso creare un nuovo dato nel seguente modo ed usare questo nuovo modello; l'altra soluzione è togliere coarse completamente e lasciare solo fine, ma secondo me non ha troppo senso.

The new column will contain the sum of coarse and 2×fine.

```
## 
## Call:
## lm(formula = rate ~ rain + coarse + fine, data = df)
## 
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.25732 -0.37283 -0.04774  0.34202  0.99202 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 20.560257   0.980962  20.959  < 2e-16 ***
## rain         0.006132   0.001024   5.989 6.54e-08 ***
## coarse       0.386020   0.047438   8.137 6.01e-12 ***
## fine         0.194142   0.045713   4.247 6.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5193 on 76 degrees of freedom
## Multiple R-squared:  0.618,  Adjusted R-squared:  0.6029 
## F-statistic: 40.98 on 3 and 76 DF,  p-value: 7.312e-16
```

```
##  (Intercept)         rain       coarse         fine 
## 20.560257374  0.006131644  0.386020255  0.194141815
```

Let us check this w.r.t. the other models:

```
## Analysis of Variance Table
## 
## Model 1: rate ~ rain + hardness + coarse + fine
## Model 2: rate ~ rain + coarse + fine
## Model 3: rate ~ rain + coarse + fine
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     75 20.478                           
## 2     76 20.492 -1 -0.014387 0.0527 0.8191
## 3     76 20.492  0  0.000000
```

This new model is a bit different from the others.

Attempt 2

We will remove coarse.

```
## 
## Call:
## lm(formula = rate ~ rain + fine, data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.57964 -0.46529 -0.02245  0.44399  1.61953 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 24.712725   1.138580  21.705  < 2e-16 ***
## rain         0.005414   0.001386   3.906    2e-04 ***
## fine         0.255664   0.061270   4.173 7.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7057 on 77 degrees of freedom
## Multiple R-squared:  0.2852, Adjusted R-squared:  0.2666 
## F-statistic: 15.36 on 2 and 77 DF,  p-value: 2.437e-06
##  (Intercept)         rain         fine
## 24.712725099  0.005414001  0.255663510
```

Let us check this w.r.t. the other models:

```
## Analysis of Variance Table
##
## Model 1: rate ~ rain + hardness + coarse + fine
## Model 2: rate ~ rain + coarse + fine
## Model 3: rate ~ rain + fine
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     75 20.478
## 2     76 20.492 -1   -0.0144  0.0527    0.8191
## 3     77 38.347 -1  -17.8545 65.3916 8.168e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This new model is a bit different from the others.

Point d

```
##            1
## fit 30.54783
## lwr 30.26596
## upr 30.82971
```