Exam: 2021/06/18

Marco Scarpelli

06 giugno, 2024

Dataset exploration

```
##   gender age height distance siblings computertime exercisehours musiccds
## 1   male  21     70       90        1           25           0.0       25
## 2 female  22     68       95        2           10           6.5       55
## 3   male  27     72      800        3           10           0.0      125
## 4   male  20     70       10        2            5           0.0        0
## 5 female  19     67      280        2            4           2.0      164
## 6 female  19     65      150        8           10           7.0       50
##   playgames watchtv
## 1         6      10
## 2         2      12
## 3         5      10
## 4         5       5
## 5         0       2
## 6         0      15
```

Point A

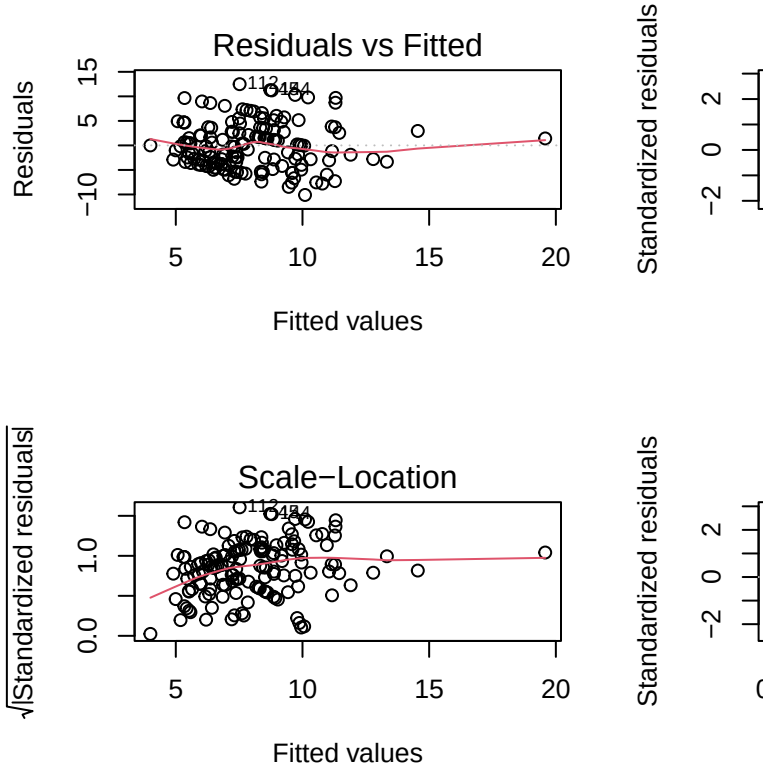
I'm a bit unsure on why they ask to explicitly encode a dependence for the gender in the intercept; if I write `1 + 1:gender + gender`, it means that there is always an interept and there is a contribution given by the gender, which is 0 if `gender`; if I just write `gender` it should be exactly the same thing. I will write it in the explicit form just to show that I complied.

I think it was a bit of a trap or bad wording and they wrote that we need to have exactly 10 parameters later in the text to ensure people would not be fooled.

The model:

```
##
## Call:
## lm(formula = watchtv ~ 1 + 1:gender + gender + age + height +
##     distance + siblings + computertime + exercisehours + musiccds +
##     playgames, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1005  -3.6037  -0.2002   3.2055  12.4808
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.4329081  9.8143851   0.655 0.513265
## gendermale     0.4381439  1.1654908   0.376 0.707546
```

```
## age            0.1948575  0.1899515   1.026 0.306768
## height        -0.1077502  0.1337050  -0.806 0.421698
## distance       0.0005002  0.0001972   2.536 0.012312 *
## siblings       0.7107674  0.2644622   2.688 0.008082 **
## computertime   0.1901959  0.0562389   3.382 0.000937 ***
## exercisehours  0.0460659  0.0978123   0.471 0.638411
## musiccds       0.0033897  0.0025794   1.314 0.190982
## playgames      0.1434152  0.1256614   1.141 0.255729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.918 on 138 degrees of freedom
## Multiple R-squared:  0.168,  Adjusted R-squared:  0.1137
## F-statistic: 3.096 on 9 and 138 DF,  p-value: 0.002037
```



Residuals; we want them to be Gaussian and homoscedastic.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fm)
## W = 0.97638, p-value = 0.01178
```

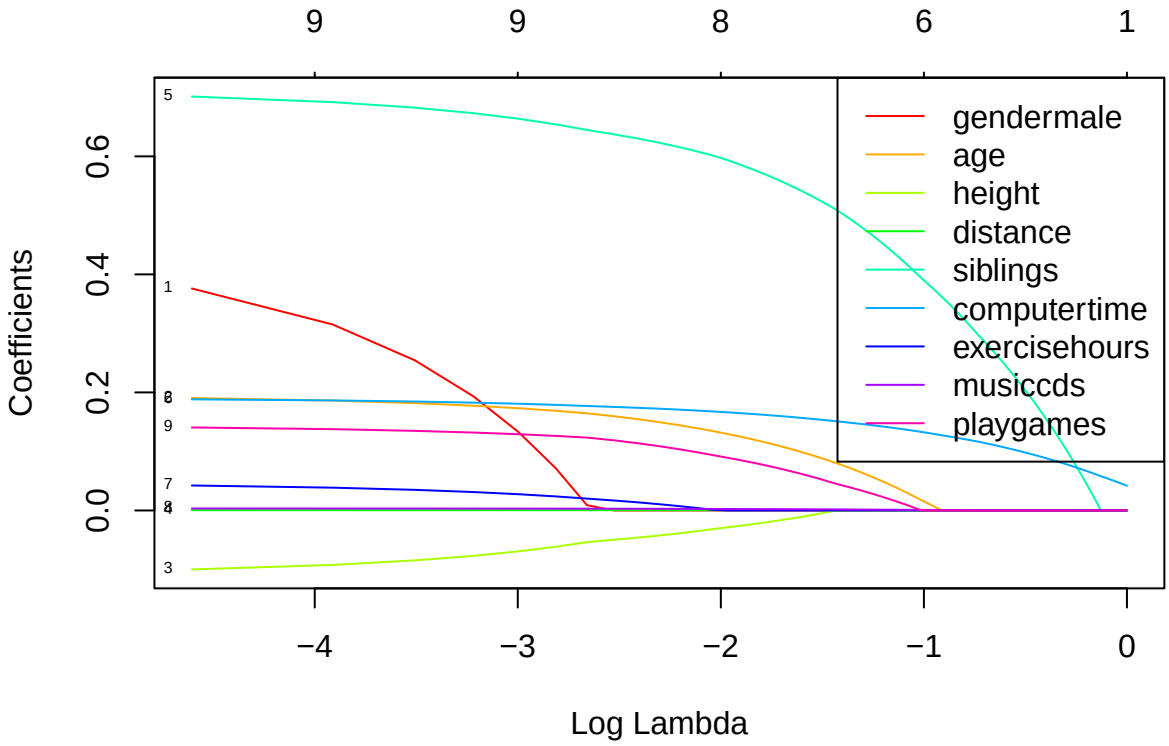
The test for Gaussianity fails; furthermore, we can see that the residuals exhibit a bit of a pattern and one of them exceeds Cook's distance.

Point B

We report the coefficients:

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## gendermale    .
## age           0.0494139480
## height        .
## distance      0.0003598352
## siblings      0.4526208912
## computertime  0.1422680784
## exercisehours .
## musiccds      0.0012825007
## playgames     0.0230406929
```

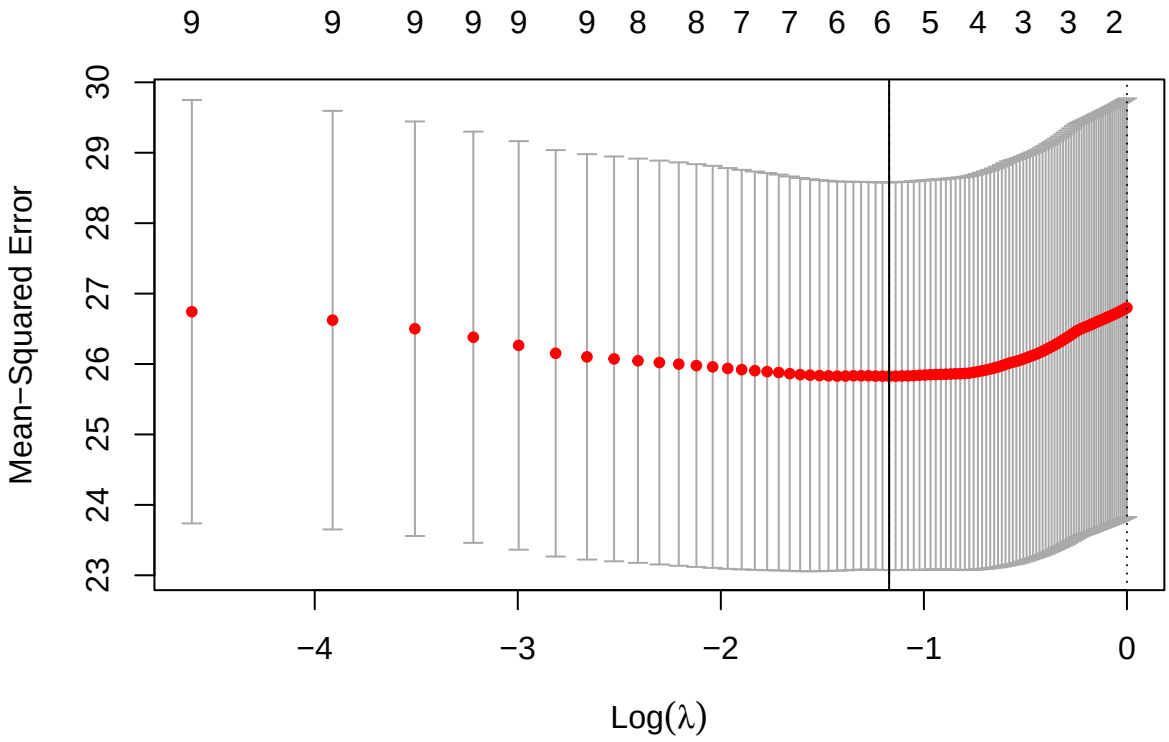
Point C



We report the best λ and the optimal λ, together with a plot to better understand the result. The black vertical linw is the best λ.

```
## [1] 0.31
```

```
## [1] 1
```



We now use the optimal λ and report the coefficients:

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## gendermale     .
## age            .
## height         .
## distance       .
## siblings       .
## computertime   0.04191207
## exercisehours .
## musiccds       .
## playgames      .
```

Point D

```
##            s1
## [1,] 7.963267
```