Applied Statistical Methods - Solution 4

Peter von Rohr

2022-03-16

Problem 1: Overfitting

Use the extended dataset on `Body Weight` of animals and fit all the variables and the factor breed. Compare the result with a regression that uses only `Breast Circumference` or with the linear model that only uses the factor `Breed`. The data set is available from: https://charlotte-ngs.github.io/asmss2022/data/asm_bw_flem.csv

Solution

- Read the data

```
s_ex04p01_data_path <- "https://charlotte-ngs.github.io/asmss2022/data/asm_bw_flem.csv"
tbl_ex04p01_data <- readr::read_csv(file = s_ex04p01_data_path)
```

```
## Rows: 10 Columns: 6
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Breed
## dbl (5): Animal, Breast Circumference, Body Weight, BCS, HEI
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

- Fit the full model

```
lm_ex04p01_full <- lm(formula = `Body Weight` ~ `Breast Circumference` + BCS + HEI + Breed, data = tbl_
summary(lm_ex04p01_full)
```

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference` + BCS + HEI +
##     Breed, data = tbl_ex04p01_data)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
##  1.8327 -0.5208  2.8604 -1.3120 -5.5552  2.6947  5.2055 -7.2432 -5.7525  7.7902
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -859.4523   513.6852  -1.673   0.1696
## `Breast Circumference`    7.1560     2.7705   2.583   0.0611 .
## BCS                       9.9056     3.8258   2.589   0.0607 .
## HEI                       0.1220     0.1822   0.669   0.5399
## BreedLimousin            13.5466    15.5227   0.873   0.4321
```

```
## BreedSimmental           -3.8614    10.1592  -0.380   0.7232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.5 on 4 degrees of freedom
## Multiple R-squared:  0.9596, Adjusted R-squared:  0.909
## F-statistic: 18.98 on 5 and 4 DF,  p-value: 0.006868
```

- Fit the model with only `Breast Circumference`

```
lm_ex04p01_bwbc <- lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_ex04p01_data)
summary(lm_ex04p01_bwbc)
```

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_ex04p01_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3941  -6.5525  -0.0673   9.3707  13.2594
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1065.115    255.483  -4.169 0.003126 **
## `Breast Circumference`     8.673      1.420   6.108 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 8 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8014
## F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

- Fit only the model with the factor `Breed`

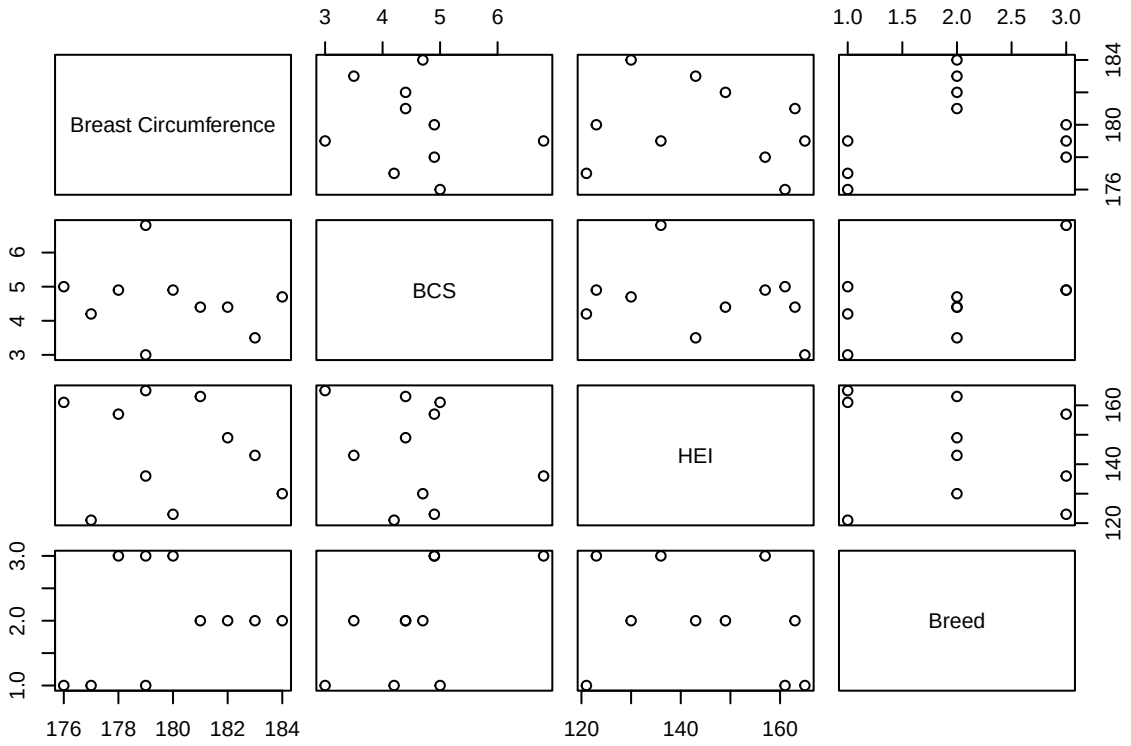
```
lm_ex04p01_bwbreed <- lm(formula = `Body Weight` ~ Breed, data = tbl_ex04p01_data)
summary(lm_ex04p01_bwbreed)
```

```
##
## Call:
## lm(formula = `Body Weight` ~ Breed, data = tbl_ex04p01_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0000  -7.5000  -0.1667   2.7500  21.0000
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     468.000      6.097  76.758 1.68e-11 ***
## BreedLimousin    52.000      8.066   6.447 0.000351 ***
## BreedSimmental   21.333      8.623   2.474 0.042575 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.56 on 7 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8196
## F-statistic: 21.44 on 2 and 7 DF,  p-value: 0.001035
```

The comparison of the models shows that the full model does not produce a better model fit. The reason for this is that the explanatory variables in the full model are correlated among each other. As a result of this correlation structure, the same information is contained in different variables and as a result the single variables do not contribute a substantial amount to the explanation of the variation in the response variable.

The correlation structure amoung the different variables can be visualized via a so called `pairs` plot.

```
tbl_ex04p01_data$Breed <- as.factor(tbl_ex04p01_data$Breed)
pairs(formula = ~ `Breast Circumference` + BCS + HEI + Breed , data = tbl_ex04p01_data)
```



From this plot, we can clearly see that `Breast Circumference` and `Breed` are correlated. If we switch levels 2 and 3 of the breeds, then we can see the relationship between `Breast Circumference` and `Breed` even better.

Problem 2: Plotting

The first step before doing any analysis should always be to plot the data which helps to visualise the internal structure of a dataset. A very instructive plot is the so-called `pairs`-plot. This plot can be done using the function `pairs()`. The task of this problem is to create a `pairs`-plot for the extended dataset on `Body Weight` of animals. The input to the function `pairs()` must be all numeric. This means that the column containing the `Breed` in our dataset must be converted to a datatype called `factor`. This can be done using the function `as.factor()`.

Results of linear models can also be plotted. In such plots, we are mainly interested in the behavior of the residuals. Hence, fit a linear regression model between `Body Weight` and `Breast Circumference` and plot the resulting linear model object.

Solution

- Read the dataset

```
s_ex04p02_data_path <- "https://charlotte-ngs.github.io/asmss2022/data/asm_bw_flem.csv"
tbl_ex04p02_data <- readr::read_csv(file = s_ex04p02_data_path)
```

```
## Rows: 10 Columns: 6
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Breed
## dbl (5): Animal, Breast Circumference, Body Weight, BCS, HEI
```

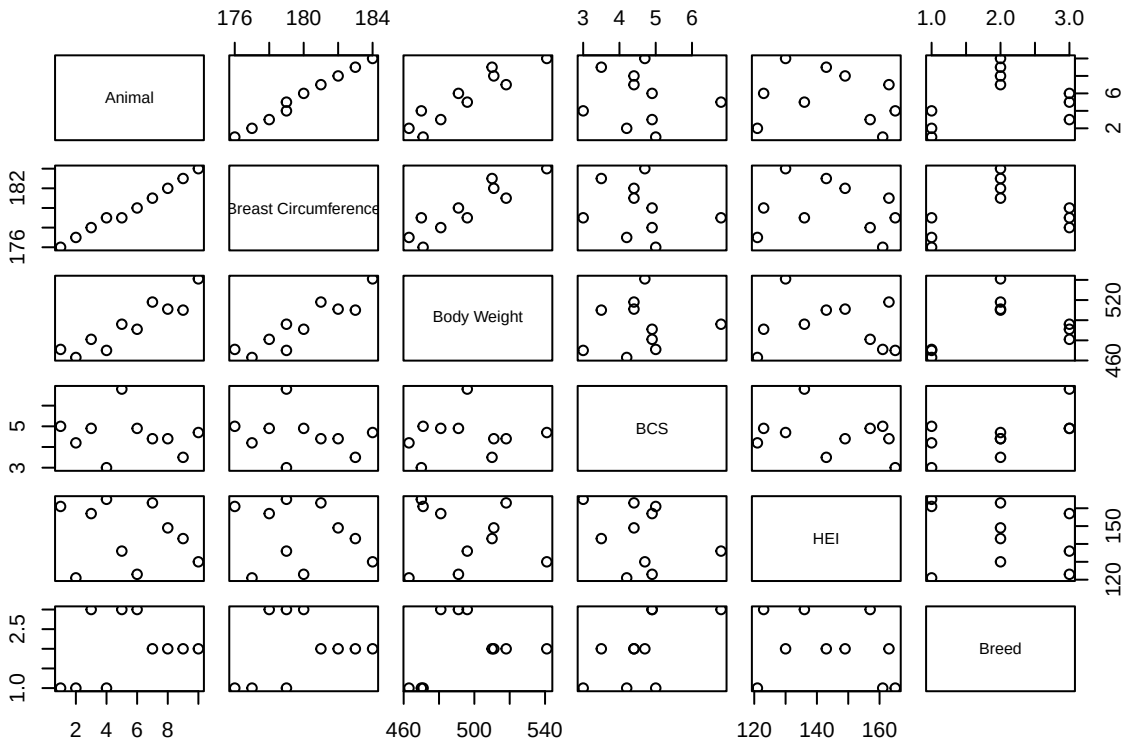
```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

- Convert the breed column to a factor

```
tbl_ex04p02_data$Breed <- as.factor(tbl_ex04p02_data$Breed)
```

- Create a `pairs`-plot

```
pairs(tbl_ex04p02_data)
```



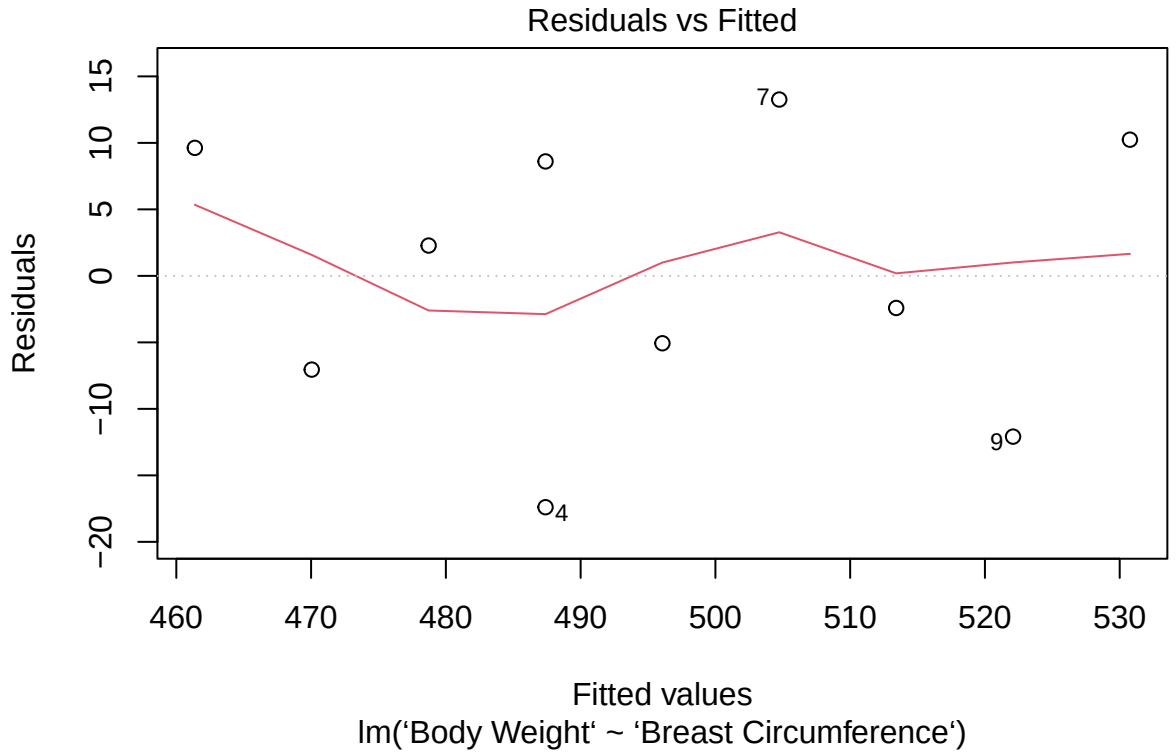
The above matrix of scatterplots shows relationships between pairs of variables.

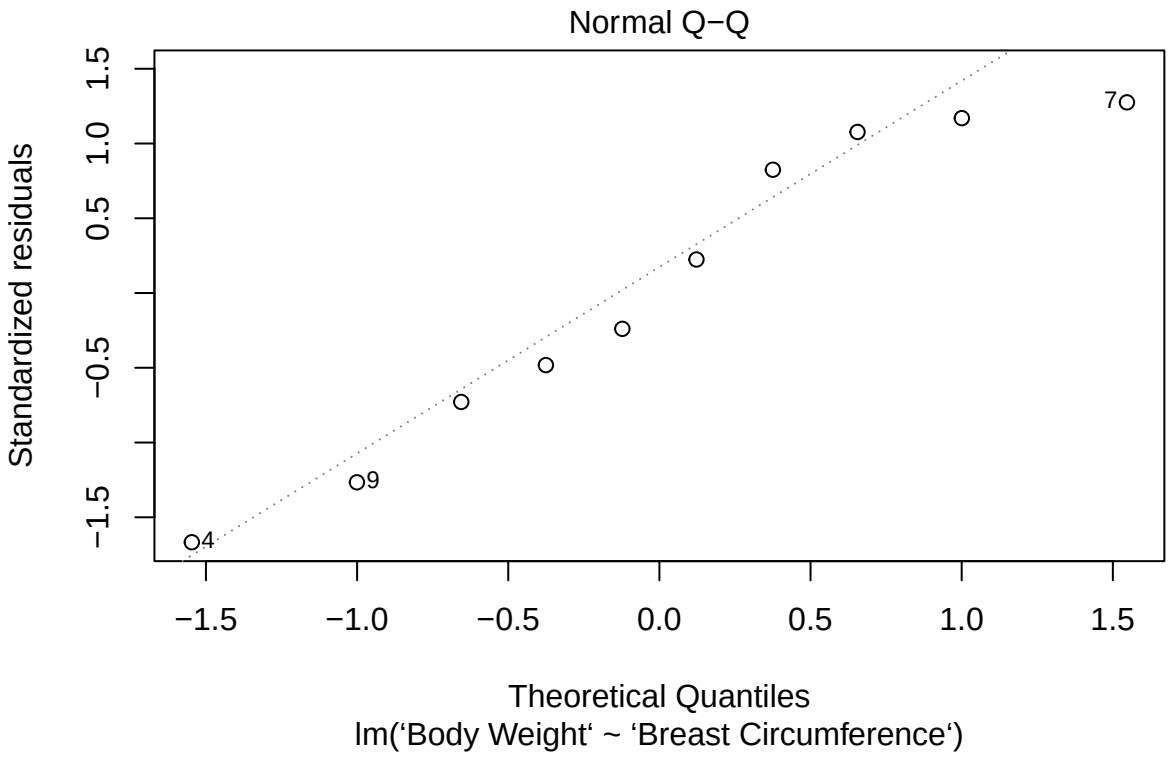
- Fit the linear regression model

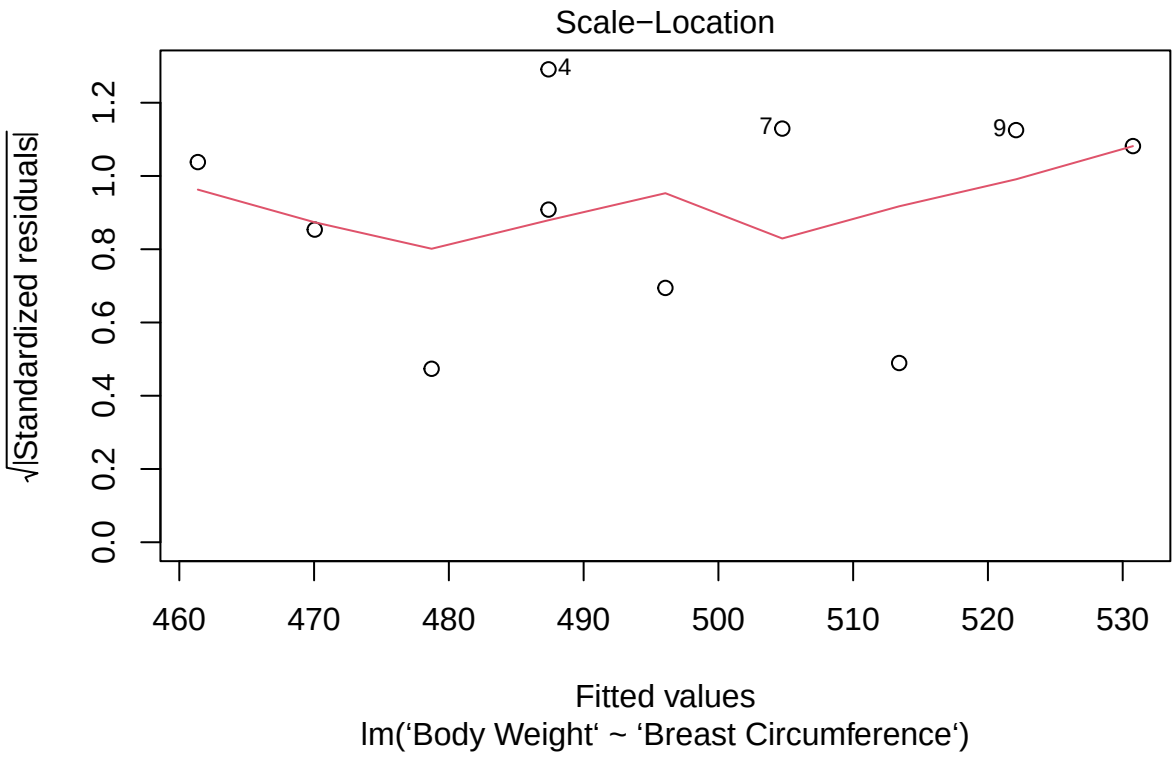
```
lm_ex04p02 <- lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_ex04p02_data)
```

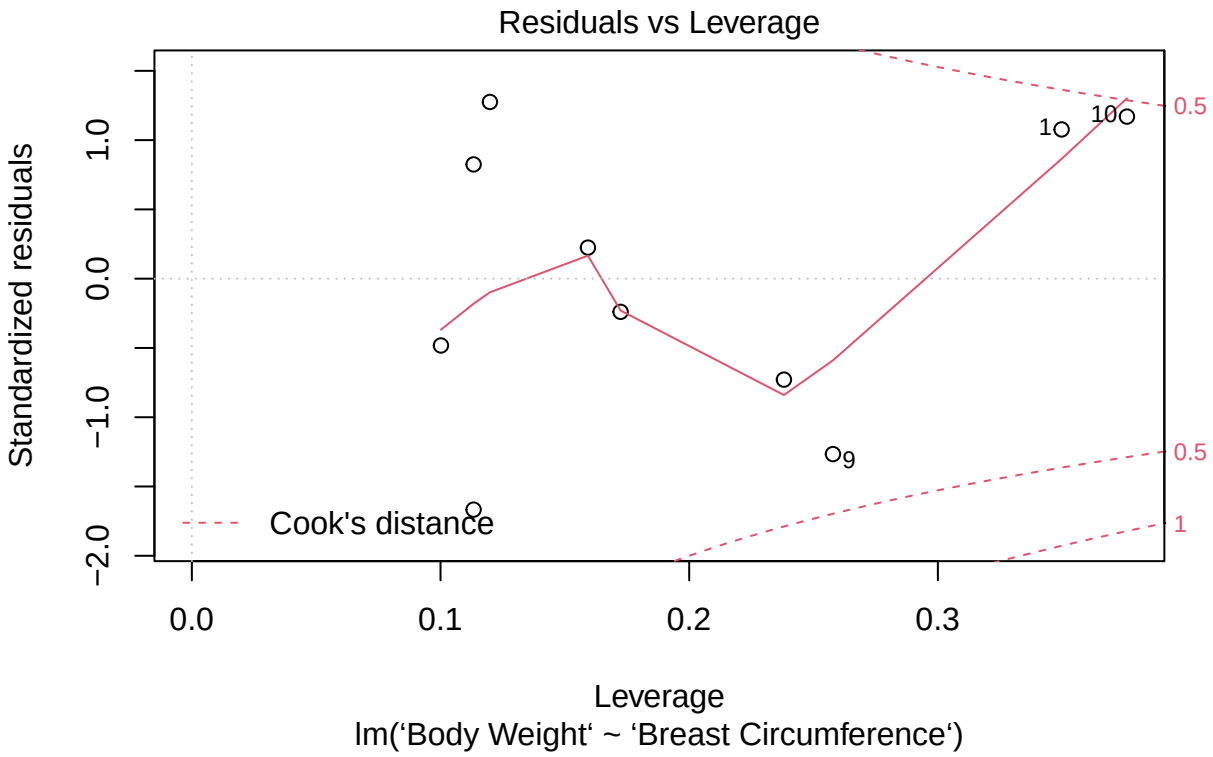
- Plot the result

```
plot(lm_ex04p02)
```




Normal Q−Q
Standardized residuals
1.5
1.0
0.5
−0.5
−1.5
7
9
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(‘Body Weight‘ ~ ‘Breast Circumference‘)


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
460
470
480
490
500
510
520
530
4
7
9
Fitted values
lm(‘Body Weight‘ ~ ‘Breast Circumference‘)



For the behavior of the resiudals, we are focusing on the first two plots. The first plot shows whether there is a dependence pattern between the residuals and the fitted values. For this plot a random pattern is desired. The second plot shows a QQ-plot of the residuals. This plot shows any deviation of the numeric distribution of the residuals from the normal distribution.