Modern Statistics and Big Data Analysis - Exam - 15 January 2024

Rules: You have 180 minutes (3 hours) time. Please send your solutions via email to

`christian.hennig@unibo.it`

before the official exam ending time (deadline), which will be announced by me when I have made the exam visible. Solutions are no longer accepted 10 minutes or more after the deadline. Late solutions that come in before 10 minutes after the deadline will normally incur a penalty loss of marks.

Only pdf, Word, and plain ASCII text/R-files are accepted. You are encouraged to type all your solutions in plain text. Handwritten solutions, given to me in person before the deadline, are accepted for the theoretical question 2 only. All other solutions need to be typed and sent by email.

Please send only one email with solutions. If you send more than one, I will only keep the last one, i.e., all earlier sent material will be lost.

Please submit only one file that has R-code in it. Please submit only one file that has typed answers in it (this can be the same as the R-file but doesn't have to be the same). You do not have to make things especially beautiful by for example using R markdown (if you use that, you do it at your own risk in case things don't work as you want).

Make sure that filenames and/or titles for graphs are chosen so that in your answers it is clear what refers to what graph. If you want to integrate the graphs in a Word or pdf file, you can do that. *In particular, make sure that you submit the graphs to which you refer in your text; don't expect me to produce the graphs myself from your code. But also comment on the graphs you submit; you need to show that you understand the graph and what can be learnt from it.*

Do not only submit R-code but also R-output (such as the clustering vector), as far as this is needed to understand and support your answers. Again, don't expect me to run your code.

Please make sure your name is on all files and all pages of your handwritten answers.

You can use all materials and the internet. You are not allowed to communicate with others. Any use of websites or services that allow an exchange of messages is strictly forbidden. Use of chatbots such as ChatGPT and more generally AI systems is strictly forbidden. There is zero tolerance. If I find any evidence that one of these rules has been broken, the exam is normally failed with a mark of zero.

Marking: I will mark on a percentage scale (1-100) and then transform results to the range 1-30. If I find out from your solutions that the exam was too difficult, I may take this into account for the transformation.
Question 1 carries 40% of the overall marks, Question 2 carries 10%, Question 3 carries 20%. The remaining 30% come from the literature question already submitted.

1. Analyse the “`Byar.dat`” data set.

Background: On the Virtuale page of the course you can find a data set named `Byar.dat`. The data set can be read into R as follows:

```
Byar <- read.table("Byar.dat")
```

The dataset gives information about 475 prostate cancer patients. There are 8 variables, mostly with self explanatory names:

Age Age of patient.

Weight Weight of patient in kg.

Systolic.Blood.pressure Systolic blood pressure of patient in units of 10.

Diastolic.blood.pressure Diastolic blood pressure of patient in units of 10.

Serum.haemoglobin Serum haemoglobin levels of patient measured in g/100ml.

Size.of.primary.tumour Estimated size of the patient’s primary tumour in centimeters squared.

Index.of.tumour.stage Combined index of tumour stage and histolic grade of the patient, the higher the more advanced the tumor is.

Serum.prostatic.acid.phosphatase Serum prostatic acid phosphatase levels of the patient in King-Armstong units.

The aim of analysis is to find and investigate potentially different types of prostate cancer. This should be done based on the variables that characterise the cancer, not `Age` and `Weight`. This means that you are expected to cluster the patients based on the variables 3-8 only, i.e., `Byarc <- Byar[,3:8]`.

For some commands you may want to use the matrix version:

```
mByar <- as.matrix(Byarc)
```

`Age` and `Weight` are provided because for the interpretation of clusters it would be interesting to know whether age and weight distribution between clusters are clearly different.

What is expected: Produce and submit at least two clusterings of the patients. Your marks will not necessarily improve if you submit more; the number of clusterings is not a marking criterion. Explain why you chose the specific clustering methods and motivate methodological decisions (such as how you choose the number of clusters).

At least one of your clusterings should only use methodology that was introduced in the course (obviously you are not expected to use any methodology that was not introduced in the course, I just won’t stop you if you want to do that; actually I advise against it because experience shows that attempts to do something far away from the course material often cost a lot of time and go wrong).

Produce at least one visualisation each for at least two clusterings.

Compare the clusterings and comment on how meaningful and useful you think they are. Select one clustering that you prefer.

Interpret the clusters (you can use all given variables and your visualisations).

Submit all R-code that you're using, with appropriate comments/explanations and output necessary to verify and understand your conclusions.

There is no single correct or best solution that I have in mind and want to see here. I'm very open to your suggestions. If you have doubts about the one you are proposing, please write these down. It's more important to have something that appropriately reflects the data than something that looks "strong" if in fact it isn't. (Note that true grouping information actually exists, but is not provided to you. I may compare your solutions graphically to the true grouping, but I will not use similarity as a marking criterion.)

2. **(a, 6%)** Consider knots at $s_1 = 1,\ s_2 = 2,\ s_3 = 3$. Define a B-spline B of order $d = 2$ so that $B(1) = 0,\ B(2) = 1,\ B(3) = 0,\ B(x) = 0$ for $x \leq 1$ and $x \geq 3$, but $B(x) > 0$ for $x \in (1, 3)$. This means that you are asked to specify the first order polynomials between any two of the knots that define B.

   (b, 4%) Consider a p-dimensional B-spline basis $\psi_1, \ldots, \psi_p$ of order $d = 2$ between knots $s_1, \ldots, s_p$. This is the basis of a p-dimensional vector space of functions V. All functions $X \in V$ can be written as

$$X(t) = \sum_{j=1}^{p} \gamma_j \psi_j(t).$$

   Give arguments why all $X \in V$ are continuous and piecewise linear.

3. In the Appendix for this question there is some R-code for an output from an analysis of a regression data set. The data contains information about 1201 sales of individual residential property in Ames, Iowa (USA) between 2006 and 2010.

   The task here is to understand the factors contributing to and predicting the house price (variable `SalePrice`, in US Dollar).

   The explanatory variables are:

   LotFrontage Linear feet of street connected to property,

   LotArea Lot size in square feet,

   OverallQual Overall material and finish quality rating,

   OverallCond Overall condition rating,

   YearBuilt Year in which property was built,

   TotalBsmtSF Total square feet of basement area,

   LowQualFinSF Low quality finished square feet (all floors),

   GrLivArea Above ground living area square feet.

   For this task I ran Least Squares and MM-regression estimators.

   (a), 4% Looking at both LS- and MM-regression results, which are the variables that look important for predicting `SalePrice`? Comment on how strong the differences between the two regression estimators are.

   (b), 4% Which of the two regression fits do you prefer and why?

   (c), 4% Are there regression outliers, good and/or bad leverage points in the data? How can you see this?

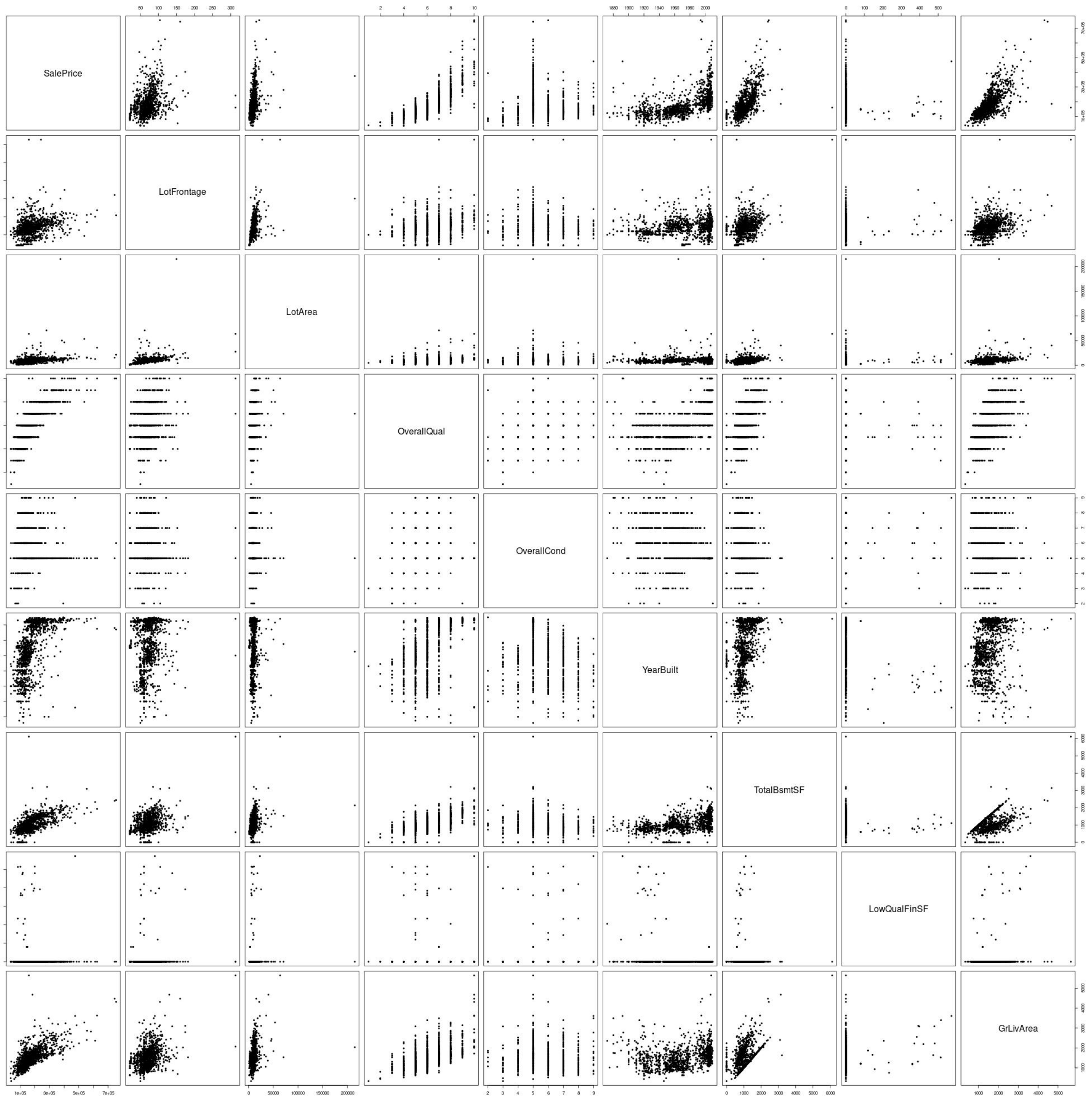
   (d), 4% The last diagnostic plot for the MM-regression (last plot in the Appendix) plots robustness weights against fitted values (the plotted numbers are observation numbers). In the course I have plotted the robustness weights against the observation number, but here the fitted values are more interesting. Comment on how robustness weights are related to fitted values, how this corresponds to what can be seen in the other plots, and what this means for the model fit.

   (e), 4% Comment on potential shortcomings of both regression fits. How do you think could an analyst try to improve the model?

Appendix for Question 3

Overview of the dataset

```
library(robustbase)
# The data are called housep
pairs(housep)
> str(housep)
'data.frame': 1201 obs. of  9 variables:
 $ SalePrice   : int  208500 181500 223500 140000 250000 143000 307000 129900 118000 129500 ..
 $ LotFrontage : int  65 80 68 60 84 85 75 51 50 70 ...
 $ LotArea     : int  8450 9600 11250 9550 14260 14115 10084 6120 7420 11200 ...
 $ OverallQual : int  7 6 7 7 8 5 8 7 5 5 ...
 $ OverallCond : int  5 8 5 5 5 5 5 5 6 5 ...
 $ YearBuilt   : int  2003 1976 2001 1915 2000 1993 2004 1931 1939 1965 ...
 $ TotalBsmtSF : int  856 1262 920 756 1145 796 1686 952 991 1040 ...
 $ LowQualFinSF: int  0 0 0 0 0 0 0 0 0 0 ...
 $ GrLivArea   : int  1710 1262 1786 1717 2198 1362 1694 1774 1077 1040 ...
```


SalePrice
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
TotalBsmtSF
LowQualFinSF
GrLivArea
50 100 150 200 250 300
2 4 6 8 10
1880 1900 1920 1940 1960 1980 2000
0 100 200 300 400 500
1e+05 3e+05 5e+05 7e+05
0 50000 100000 150000 200000
2 3 4 5 6 7 8 9
0 1000 2000 3000 4000 5000 6000
1000 2000 3000 4000 5000

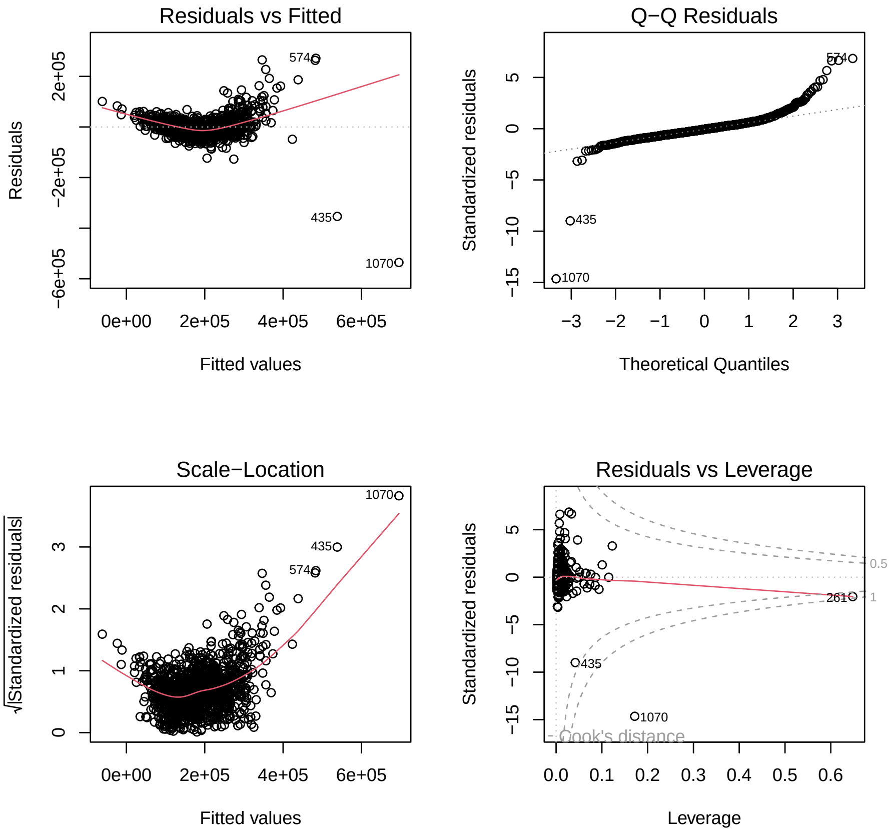
Least Squares regression

```
lmhouse <- lm(SalePrice˜.,data=housep)
summary(lmhouse)

(...)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.180e+06  1.033e+05 -11.424  < 2e-16 ***
LotFrontage   2.375e+01  5.688e+01   0.418    0.676
LotArea       8.288e-01  1.662e-01   4.987 7.03e-07 ***
OverallQual   2.251e+04  1.357e+03  16.593  < 2e-16 ***
OverallCond   6.592e+03  1.187e+03   5.555 3.42e-08 ***
YearBuilt     5.410e+02  5.278e+01  10.252  < 2e-16 ***
TotalBsmtSF   2.706e+01  3.434e+00   7.881 7.29e-15 ***
LowQualFinSF -4.086e+01  2.355e+01  -1.735    0.083 .
GrLivArea     5.412e+01  3.128e+00  17.301  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40180 on 1192 degrees of freedom
Multiple R-squared:  0.7694,Adjusted R-squared:  0.7679
F-statistic: 497.2 on 8 and 1192 DF,  p-value: < 2.2e-16
```

```
# Diagnostic plots
par(mfrow=c(2,2))
plot(lmhouse,ask=FALSE)
```



Robust regression

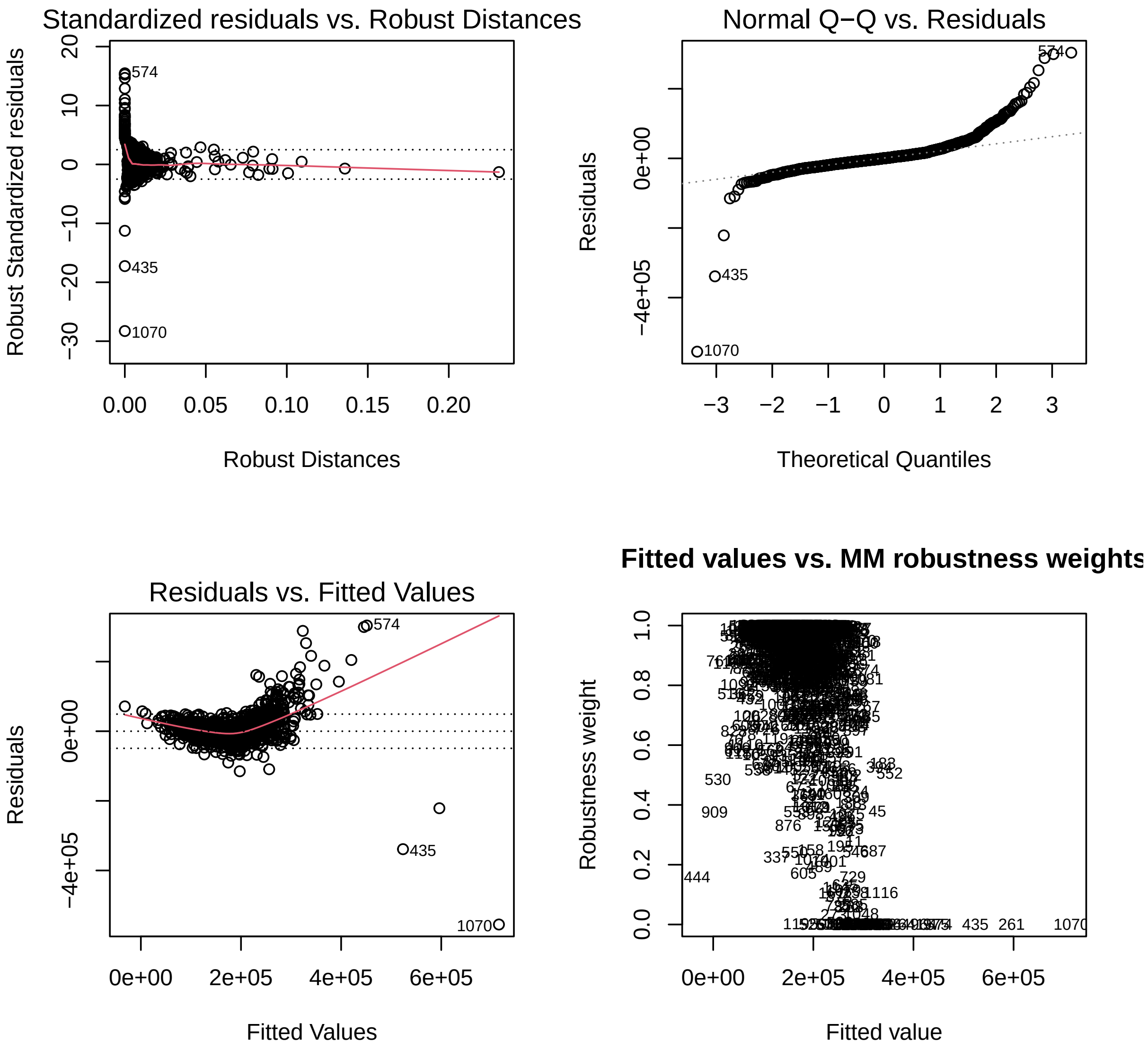
```
lmrhouse <- lmrob(SalePrice˜.,data=housep)
summary(lmrhouse)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.252e+06  5.901e+04 -21.207  < 2e-16 ***
LotFrontage   7.739e+01  4.738e+01   1.634   0.1026
LotArea       1.710e+00  3.133e-01   5.457 5.88e-08 ***
OverallQual   1.448e+04  1.157e+03  12.519  < 2e-16 ***
OverallCond   7.686e+03  7.167e+02  10.724  < 2e-16 ***
YearBuilt     5.925e+02  3.082e+01  19.222  < 2e-16 ***
TotalBsmtSF   2.982e+01  2.659e+00  11.214  < 2e-16 ***
LowQualFinSF -5.373e+01  2.057e+01  -2.613   0.0091 **
GrLivArea     4.928e+01  2.547e+00  19.346  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 19630
Multiple R-squared:  0.8677,Adjusted R-squared:  0.8668
Convergence in 20 IRWLS iterations

Robustness weights:
 43 observations c(50,96,126,133,146,152,190,232,261,290,310,320,364,390,398,429,
435,439,486,492,526,537,555,566,572,574,641,645,664,669,747,818,862,944,964,974,
975,1017,1030,1070,1090,1141,1182)
 are outliers with |weight| <= 5.6e-05 ( < 8.3e-05);
 98 weights are ˜= 1. The remaining 1060 ones are summarized as
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001809 0.842300 0.949800 0.867300 0.985900 0.999000
```

```
# Diagnostic plots
plot(lmrhouse,ask=FALSE,which=c(1,2,4))
plot(lmrhouse$fitted,lmrhouse$rweights,type="n",main="Fitted values vs. MM robustness weights"
text(lmrhouse$fitted,lmrhouse$rweights,cex=0.7)
```


Standardized residuals vs. Robust Distances
Robust Standardized residuals
Robust Distances
Normal Q−Q vs. Residuals
Residuals
0e+00
−4e+05
Theoretical Quantiles
Residuals vs. Fitted Values
Residuals
0e+00
−4e+05
0e+00
2e+05
4e+05
6e+05
Fitted Values
Fitted values vs. MM robustness weights
Robustness weight
0e+00
2e+05
4e+05
6e+05
Fitted value