Robust Statistics (2600WETROS)
Project June 2022-2023

1 General

- This project is part of the exam.
- The intention is to use methods from the course, in addition to this one can explore other techniques as well.
- Every student should write a report written with enough care. Interesting figures or when figures are explicitly asked, may be put in the the text. Other graphs should be part of an appendix.
- For every question the corresponding code has to be submitted. Make sure it structured nicely, documented and may be executed from start to finish. Your guideline for this are the solution scripts of the exercise sessions. Including the scripts is enough, there is no need to include the code in the written report.
- The report, together with the R-files has to be submitted via Blackboard no later than two weeks before the exam. Exact procedure will be communicated via Blackboard.

2 Datasets

The following data sets will be part of this assignment. A short description is given below.

2.1 Food data

The food dataset is derived from the USDA National Nutrient Database. It contains variables expressing how much of certain substance (eg. fat or copper) is contained in the food. For your information a description of the good and the food group it belongs to is provided. These two variables should not be included in the analysis, but can be used to interpret results.

2.2 Topgear data

The Topgear consists of 242 cars that appeared in the British show Topgear. For each car the following variables are available

Variable	Description
`Maker`	the brand of the car
`Model`	the model of the car
`Price`	the price of the car (in Britsh pounds)
`Displacement`	cilinder volume (in cc)
`BHP`	Power of the engine (in bhp)
`Torque`	Torque of the car (in lb/ft)
`Acceleration`	number of seconds to 62 miles per hour
`MPG`	fuel consumption (in miles per gallon)

3 Questions

3.1 Question 1

a) Write a function that calculates the M-estimator for location using the IWLS (Iteratively Reweighted Least Squares, also known as IRWLS) algorithm. Do this for both Huber's rho function ($b = 1.5$) and Tukey's bisquare function ($c = 4.68$). Use the median and `MAD` as initial estimators. For this write several aid functions that you can in the algorithm and to structure your code:

- A function implementing the Huber's rho and Tukey's bisquare function
- A function to calculate the weights in the IRLS step

b) Generate a symmetric sample from the standard normal distribution of size n = 100. Calculate and plot the sensitivity curve for the 2 estimators based on this sample. For this, add the observations z = $z = (-4, -3.99, \ldots, 4)$ to the fixed generated data sample one by one. Discuss your findings of the results.

c) Repeat b) for a sample of size $n > 10^6$ (or what your computer can handle in terms of computing power). What curve are we approaching? Discuss!

3.2 Question 2

For the food data:

a) Consider the variable *Coppermcg*. Discuss whether this data might stem from a normal distribution. If not, does transforming the data results in data that is closer to a normal distribution. What about robustness aspects? Discuss!

b) Perform a robust transformation using Yeo-Johnson power transformation for all the variables. You may ignore possible warnings given by the routine. Investigate if robust dimension reduction techniques can be used to (interpretably) represent the data in a lower dimension. Consider the ROBPCA procedure. What are your conclusions? Is it appropriate to use ROBPCA for this data set? Discuss in detail.

3.3 Question 3

For the Topgear data:

a) Perform an exploratory analysis of the TopGear data. Use graphical representations and statistics to gain insights into the data. Report and illustrate interesting findings.

b) Conduct a classic regression analysis that predicts the price of a car based on the other numerical variables in the data. Investigate whether transforming the response variable yields a better model. Interpret the regression coefficients and check the model assumptions.

c) Create diagnostic plots and check for outliers. What type of outliers are they? Can you explain them based on the information in the dataset (model/brand)?

d) Repeat b) and c), but this time using robust estimators. Compare the results and describe your conclusions.

3.4 Question 4

Consider the warpbreaks data set from the R library datasets.

a) Consider a Possion glm model to predict the number of breaks given the wool and tension variables. How do you encode the variables? What is the effect of the different levels of tensions on the mean number of breaks? Discuss!

b) What can you say on over- or underdispersion? model. Is a different type of glm model more appropriate? Explain your answer.

c) Repeat the analysis of question a), but now you a robust method. Are results in line with what you expected? Discuss.

Good Luck!