Master in Statistics and Data Science , Hasselt University

Computer Intensive Methods: Final projects (2024/2025)

Introduction

The projects

The final project in the course consists of 3 mini projects. In each project you are asked to conduct an analysis of one or more dataset(s). In some cases, to access the data, you need to install R packages. The datasets are a part of these packages. You can get more information about the data using a link provided at a description of the data.

Complete case analysis

For all projects, **only observations without missing values (complete case analysis) should be included in the analysis**. Use the R function na.omit(data) to exclude all observations with missing values.

General information

1. The projects should be done in groups of 3-4 students.
2. For each project, write a short report of maximum 12 pages with your solution for the questions.
3. Formulate clearly the hypotheses that you wish to test and/or the statistic(s) of interest.
4. Whenever relevant, formulate clearly the statistical model(s) that you use.
5. Present **clearly** the bootstrap algorithm(s) (or the re sampling algorithm(s)) that you use. In case that there is more then one re sampling method in a question, formulate all re sampling algorithms that you use.
6. Do not include R code in the report but prepare a separate appendix in which R code is listed to allquestions.
7. For each question, discuss and interpret the results.

What do you need to submit as a solution?

1. You need to submit answers to the questions. In your report focus on the questions and give clear answers. **Do not** describe the data structure in your answer.
2. Answers without a clear explanation about the results will not receive points. For example, if the question is: “Use parametric bootstrap to calculate a 95% for the mean” and you answer will be: “the C.I is (10.55-10.79)” you will **not** receive points.

When do you need to submit the solution?

- Submission date: January 2025 (exact date TBA).

- Time: 17:00.

How to submit the solution?

The solutions should be sent by email (one email per group) to ziv.shkedy@uhasselt.be. An information email will be sent in early January 2025.

Oral exam: an open book exam

1. The oral exam will be an **individual** oral exam of (approx.) 20-30 minutes per students. The questions in the exam will be related to your homeworks and selected topics from the course.
2. The **individual** oral exam will take place in January 2025. More information about the exam and the time schedule for the exam will be sent by email in a later stage.
3. The oral exam is an open book exam, you can use any material that you find useful to solve the questions (this include course slides, online books, websites, etc).
4. You are **NOT** allowed to discuss the exam and projects with students from other groups. Do not share your code with students from other groups and do not share your solutions with students from other groups.
5. Bring your solutions to the projects with you to the exam.

Project 1

In this project, in all questions, we focused on the `nassCDS` data which is a US data from police-reported car crashes (1997-2002) in which there is a harmful event (people or property). Data are restricted to front-seat occupants, include only a subset of the variables recorded. More information about the dataset can be found using the following link: https://www.rdocumentation.org/packages/DAAG/versions/1.22/topics/nassCDS. The data is a part of the `DAAG` R package. To get an access to the data you first need to install the package. The list of variables names is shown below.

```
library("DAAG")
data(nassCDS)
names(nassCDS)

##  [1] "dvcat"       "weight"      "dead"        "airbag"      "seatbelt"
##  [6] "frontal"     "sex"         "ageOFocc"    "yearacc"     "yearVeh"
## [11] "abcat"       "occRole"     "deploy"      "injSeverity" "caseid"

dim(nassCDS)

## [1] 26217    15
```

Question 1:

Let Y_i be an indicator variable which takes the value of 1 if an occupant died in an accident (the variable `dead`) and zero otherwise and X_i be the age of occupant in years (the variable `ageOFocc`). We consider the following GLM:

$$g\big(P(Y_i = 1)\big) = \beta_0 + \beta_1 X_i.$$

1. Estimate the model using a classical GLM approach.
2. Let X_{50} be the age of occupant for which the probability to die is 0.5, i.e., $P(Y_i = 1) = 0.5$. Estimate X_{50}. Use non parametric bootstrap to estimate the distribution of X_{50} and to construct a 95% C.I. for the X_{50}.
3. For the model formulated above, estimate the OR (for a unit increased in age). Use non parametric bootstrap to construct a 95% C.I. for the OR (for a unit increased in age) using the percentile and bootstrap t interval methods, which one do you prefer for the parameter OR ?
4. We focus on the odds ratio (OR) for a unit increased in age. Use parametric bootstrap to test the null hypothesis $H_0: OR = 1$.
5. Let π_{33} be the probability of death for an occupant at age 33. Use parametric bootstrap to calculate the standard error for $\hat{\pi}_{33}$ and construct a 90% C.I. for π_{33}.

Question 2

In this question we fit a robust GLM for the model specified in Q1. Use the R package `glmRob` to fit the model.

1. Estimate the model using the R package `glmRob`.
2. Use non parametric bootstrap to estimate the SE for the intercept and slope.
3. Use the jackknife and the bootstrap procedures to estimate the bias and MSE for the intercept and slope estimated by: (1) the GLM model in Q2.1 and (2) the robust GLM model estimated in Q2.2. Which method you prefer to use for the estimation of the intercept and slope?

Question 3:

In this question we focus of the following 2 × 2 table (for the complete case analysis) for the variables `airbag` and `dead`.

```
##
##          alive  dead
##   none   11058   669
##   airbag 13825   511
```

1. Define the observation unit (X_i, Y_i) for the question.
2. Calculate the odds ratio for usage of airbag and the accident outcome (dead/alive) and construct 95% confidence interval. You can use the R function `oddsratio`. What is your conclusions ? Do you think that airbags in the car influence the accident outcome ?
3. Use parametric bootstrap to construct a construct a 95%confidence interval for the OR.
4. Use permutations test to test the hypothesis that airbags in the car DO NOT influence the accident outcome using a chi-square test for a 2 × 2 table. Compare the distribution of the chi-square test statistic in this question to the theoretical distribution of the test statistic.

Question 4:

If this question we focus on the variables `dead` (the outcome of the accident) and the gender (the variable `sex`).

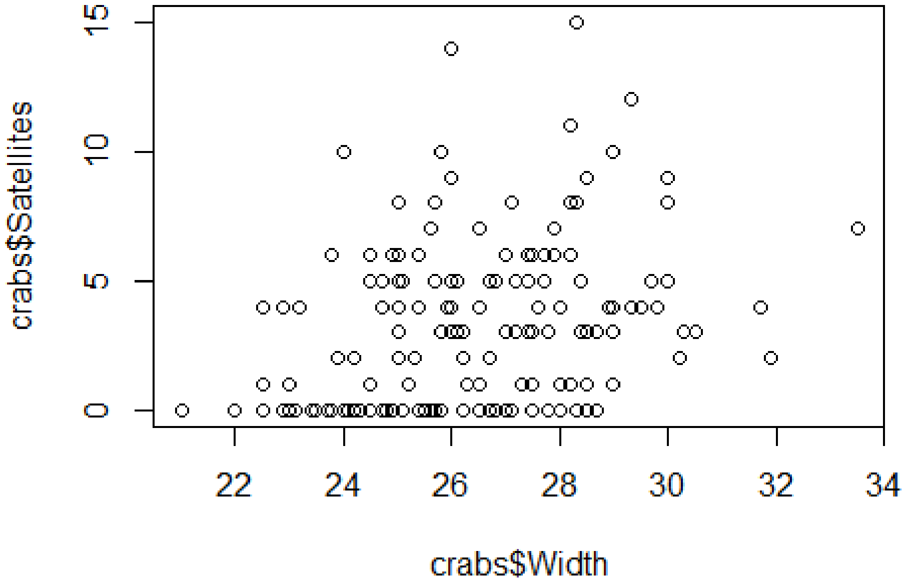
1. Estimate the proportion of male (π_M) and female (π_F) that died in the accidents.
2. Test the hypothesis that the proportion of male and female that died in an accident are equal using a classical two-samples test (use a two sided test).
3. Use parametric bootstrap to test the hypothesis that the proportion of male and female that died in an accidents are equal against a two sided alternative.
4. Use non-parametric bootstrap construct a 95% confident interval for $\pi_M - \pi_F$.

Project 2

Question 1:

In this question we use the horseshoe crab dataset. The data is available in R (`crabs`) as a part of the R package `glm2`. To get the data, install the package glm2 and use the code below to access the data.

```
library(glm2)
data(crabs)
names(crabs)

## [1] "Satellites" "Width"      "Dark"       "GoodSpine"  "Rep1"
## [6] "Rep2"

plot(crabs$Width,crabs$Satellites)
```



The dataset contains information about of 173 female horseshoe crab. You can find more details about this dataset in the book of Alen Agresti (An Introduction to Categorical Data Analysis, Section 3.3.2). The first 6 lines are given below.

```
head(crabs)

##   Satellites Width Dark GoodSpine Rep1 Rep2
## 1          8  28.3   no        no    2    2
## 2          0  22.5  yes        no    4    5
## 3          9  26.0   no       yes    5    6
## 4          0  24.8  yes        no    6    6
## 5          4  26.0  yes        no    6    8
## 6          0  23.8   no        no    8    8
```

Each female horseshoe crab in the study had a male crab attached to her nest. The study investigated factors that affect whether the female crab had any other males, **called satellites**, residing nearby her. The response outcome for each female crab is her number of satellites (`Satellites`). In this question, possible explanatory variables are the female crab's shell width (`Width`), which is a summary of her size and a binary factor indicating whether the female has good spine condition (yes or no, in R: `GoodSpine`).

1. Let Y_i be the number of satellites,we assume that $Y_i \sim \text{Poisson}(\mu_i)$ where μ_i denotes the expected number of satellites for the i'th female crab. We consider the following linear predictor:
$$g(\mu_i) = \beta_0 + \beta_1 \times \text{Width}_i + \beta_2 \times \text{GoodSpine}_i.$$
Here, $g(\quad)$ is the link function. Formulate an appropriate model for the number satellites. Fit the model and use the likelihood ratio test in order to test the null hypothesis $H_0\colon \beta_2 = 0$ against a two sided alternative.
2. Use parametric and non parametric bootstrap to test the null hypothesis in Q1.1. Compare the distribution of the likelihood ratio statistic obtained for the two bootstrap procedures to the theoretical distribution of the likelihood ratio test, what is you conclusion ?
3. Use permutaions test to test the null hypothesis formulated in Q1.1.

Question 2:

The data we use for this question is the sleep data. The study was conducted to show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients. The variable `extra` is the response variable , represents the increase in hours of sleep due to the treatment, and the variable `group` is the grouping factor. The data is given below.

```
extra<-c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0,
         1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,4.3)
group<-c(rep(1,10),rep(2,10))
ID<-c(1:20)
data.frame(extra,group,ID)
```

```
##    extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
## 11   1.9     2 11
## 12   0.8     2 12
## 13   1.1     2 13
## 14   0.1     2 14
## 15  -0.1     2 15
## 16   4.4     2 16
## 17   5.5     2 17
## 18   1.6     2 18
## 19   4.6     2 19
## 20   4.3     2 20
```

let μ_1 and μ_2 be the means of the first and the second treatment group, respectively. We wish to test the null hypothesis

$$H_0: \mu_1 = \mu_2,$$

against a two sided alternative.

1. Use the classical two-samples t-test for two independent samples.
2. Non parametric bootstrap test using the two-samples t-test statistic as a test statistic.
3. Let M_1 and M_2 be the sample medians of first and the second group, respectively. The test statistic t_M is given by
$$t_m = \frac{M_1 - M_2}{SM(x)}.$$
where
$$SM(x) = \Sigma_{i=1}^{10}|x_{1i} - M_1| + \Sigma_{i=1}^{10}|x_{2i} - M_2|$$
Use a parametric bootstrap to test the null hypothesis using the test statistics t_m.
4. Compare the distribution of the test statistics in Q2.2 and Q2.3.

Question 3:

We consider the following dataset with three variables and 10 observations.

```
ID<-c(1:10)
x1<-c(0.8,-1.23,1.25,-0.28,-0.03,0.61,1.43,-0.54,-0.35,-1.60)
x2<-c(0.64,-1.69,1.47,-0.14,-0.18,0.43,1.61,-0.31,-0.38,-1.82)
data.frame(ID,x1,x2)
```

```
##    ID    x1    x2
## 1   1  0.80  0.64
## 2   2 -1.23 -1.69
## 3   3  1.25  1.47
## 4   4 -0.28 -0.14
## 5   5 -0.03 -0.18
## 6   6  0.61  0.43
## 7   7  1.43  1.61
## 8   8 -0.54 -0.31
## 9   9 -0.35 -0.38
## 10 10 -1.60 -1.82
```

Note that there two observations per subject: (X_{1i}, X_{2i}) which represent a measurement of the same variable **before** and **after** a treatment. The statistic of primary interest in this question is the ratio between the means, that is

$$\hat{\theta} = \frac{\bar{X}_1}{\bar{X}_2}.$$

1. Estimate the ratio statistic.
2. Estimate the standard error of the ratio using non parametric bootstrap and Jackknife. For the bootstrap procedure use: B=10,20,50,100,250,500,1000,2500,5000,7500,10000. Which value of B you recommend to use ?
3. Construct a 95% bootstrap confidence interval for the ratio.
4. Use a bootstrap procedure to test the hull hypothesis $H_0: \theta = 1$ against a one sided alternative. **Do not** use a two-samples paired t-test for the mean difference to test the null hypothesis.

Question 4:

Consider the data in Q3, let Let μ_1 and μ_2 the mean of the subjects' first and the second measurements, respectively. Let the mean deference $\mu_d = (\mu_1 - \mu_2) = \big(E(X_{1i}) - E(X_{2i})\big)$.

1. Construct a 95% C.I for μ_d using the classical method.
2. Use non parametric bootstrap to construct a 95% C.I for μ_d. Use the percentile, bootstrap t and BCa methods to construct the C.I.
3. Test the hypothesis $H_0: \mu_d = 0$ using a non parametric bootstrap procedure.

Project 3

Question 1:

The data we use in this question is the Chicks dataset. The data contains information over an experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. use the following code to acsess the data:

```
head(chickwts)

##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

The question of primary interest is if there is a difference between the chicks weights across the diet groups. Let Y_{ij} be the weight of a chick i in diet group j.

1. Formulate a one-way ANOVA model for the problem, formulate the null hypothesis and the alternative. Test the null hypothesis using the classical F test. Formulate the test statistic and test the null hypothesis using significance level of 5%.
2. Use semi- parametric bootstrap in order to test the null hypothesis of no diet effect.
3. Use permutations test to test the null hypothesis of no diet effect.
4. Let $\theta = \mu_{sunflower} - \mu_{soybean}$ be the mean difference between the Sunflower and Soybean diet groups. Estimate θ and construct a 90% C.I. for θ using a parametric bootstrap.

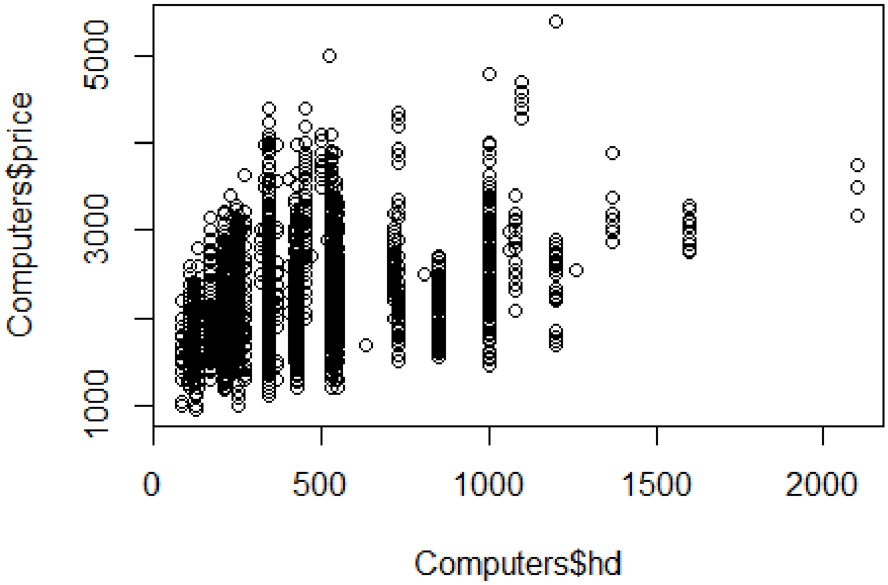
Question 2:

In this question we focused on the `Computers` dataset that can be accessed via the R package `Ecdat` . Make sure you install the package `Ecdat` in order to access the data. This data shows the prices of Personal Computers from 1993 until 1995. It contains with 6259 observations on 10 variables. Visit https://rdrr.io/cran/Ecdat/man/Computers.html to read more about the data set. Use the code below to acsess the data.

```
library(Ecdat)
data("Computers")
names(Computers)

##  [1] "price"   "speed"   "hd"      "ram"     "screen"  "cd"      "multi"  
##  [8] "premium" "ads"     "trend"

head(Computers)

##   price speed  hd ram screen cd multi premium ads trend
## 1  1499    25  80   4     14 no    no     yes  94     1
## 2  1795    33  85   2     14 no    no     yes  94     1
## 3   1595     25 170    4      15 no     no      yes  94      1
## 4   1849     25 170    8      14 no     no       no  94      1
## 5   3295     33 340   16      14 no     no      yes  94      1
## 6   3695     66 340   16      14 no     no      yes  94      1
```

let us focus on the variables “price in US dollars of 486 PCs” (the variable `price` in the dataset) and size of hard drive in MB (the variable `hd` in the dataset). Let Y_i be the price and X_i be the size of hard drive in MB.

```
##  [1] "price"   "speed"   "hd"      "ram"     "screen"  "cd"      "multi"
##  [8] "premium" "ads"     "trend"
```



We consider the following regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

1. Estimate the model using the classical OLS approach.
2. Use the estimated model to predict the price and estimate the prediction error.
3. Use a 10 fold cross validation procedure to predict the price and estimate the prediction error, compare with the results in Q2.2.
4. Use leave one out cross validation to investigate the change in the slope estimate ($\hat{\beta}_1$) as the data change and visualize this change in a graphical display.
5. Use a bootstrap procedure to construct a 95% C.I. for the predicted values (i.e., the regression line) of the model.

Question 3

In this question we use the same model formulated in Q2.

1. Use non parametric bootstrap to constract a 95% C.I for $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$.
2. Explain and illustrate how can you use a bootstrap procedure to investigate the influence of the observations for which the hard drive size is larger than 2000 MB. In your illustration, use the estimates for the $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ that were calculated in Q3.1.

Question 4

Consider a sample of 20 observations from a population with mean μ:

```
x<-c(0.68446806,-0.02596037,-0.90015774,0.72892605,-0.45612255, 0.19311847,
 -0.13297109, -0.99845382,  0.37278006, -0.20371894, -0.15468803,  0.19298230
, -0.42755534, -0.04704525,  0.15273726, 0.03655799, 0.01315016, -0.59121428,
4.50955771, 2.87272653)
length(x)

## [1] 20
```

1. Estimate μ using the mean and the median.
2. Approximate the distribution of the sample mean and the median using non parametric bootstrap with B=1000.
3. Estimate the standard error of the sample mean and the median and calculate 95% C.I for the sample mean and median using a semi parametric bootstrap.
4. Estimate the MSE for the mean and the median using jackknife, which parameter estimate you prefer to use ?.
5. Let M be the median and let $\pi_{(M<0)} = P(M < 0)$. Estimate $\pi_{(M<0)}$, Estimate the distribution of $\hat{\pi}_{(M<0)}$ and construct a 95% C.I. for $\pi_{(M<0)}$.