Assignment Classes 01-02

GH

2024-10-22

Table of Contents

## Instructions

In this exercise you will create your *first own markdown report*. You will mostly perform analyses that we have already seen in the RStudio class, but with a different data set. This exercise should help you to setup your RStudio environment properly.

Generate an R markdown report in which you perform the following analyses. Use ‘MB2-Rstudio-template.docx’ as template. (See the header above; it will automatically use this template, but you must have downloaded it into your R project folder.)

* Please fill in text whenever you see XXX in this file.
* Please fill in R code whenever you see an (empty) code chunk field beginning with:

# your R code:

*Finally, knit the markdown document and submit the knitted Word file!*

## Authors of this assignment

Please state name(s) and Matrikelnummer of all (up to 3) students who contributed to this assignment!

1. Michael Adam, 01507402
2. Aaron Paul Osburg, 12010313

##### page break

## Exercise 1

Find medical papers of your choice (enter the DOI) in which data analysis is used to …

* … answer a descriptive research question.

DOI: <https://doi.org/10.1038/s41591-024-03117-0>

Title of the paper: Disparities in air pollution attributable mortality in the US population by race/ethnicity and sociodemographic factors

Describe the research question and give an explanation why you think it is a *descriptive* question: This paper estimates the degree to which premature mortality in the USA by race/ethnicity, education, rurality and social vulnerability can be attributed to differences in exposure and susceptibility to published particulate matter (PM2.5) air pollution estimates. This research question is descriptive because it describes the association of dependent variable (mortality by race, education,…) with independent variable(s) (air pollution estimates).

* … answer a predictive research question.

DOI: <https://doi.org/10.1177/2055217319885983>

Title of the paper: Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression

Describe the research question and give an explanation why you think it is a *predictive* question: This study evaluates individual and ensemble model performance built using decision tree (DT)-based algorithms compared to logistic regression (LR) and support vector machines (SVMs) for predicting secondary progressive multiple sclerosis (SPMS) disability progression. Each of these models addresses a predictive research question because they are designed to provide an accurate prediction of an outcome (SPMS disability progression).

* … answer a causal research question.

DOI: <https://doi.org/10.1038/s41598-024-77131-0>

Title of the paper: Causal relationships between allergic and autoimmune diseases with chronic rhinosinusitis

Describe the research question and give an explanation why you think it is a *causal* question: This study investigates the causal relationships between allergic and autoimmune diseases (AR, asthma, AD, and psoriasis) and Chronic rhinosinusitis (CRS). It is a causal research question because it explains how certain variables, in this case genetic variants associated with autoimmune/allergic diseases, (causally) affect an outcome, in this case being diagnosed with CRS.

##### page break

## Exercise 2

Consider that you plan to perform a regression analysis to investigate the association of an outcome variable Y with 4 independent variables X1, X2, X3, X4. What properties of the data should be checked during an “initial data analysis”?

1. The univariate distribution of the dependent and independent variables
2. The association between the independent variables
3. Missing values and patterns of missing values

##### page break

## Exercise 3

### First, download the data set

Just execute the R chunk below.

url\_diab <- "https://hbiostat.org/data/repo/diabetes.csv"  
diabetes <- read.csv(file=url\_diab)  
head(diabetes)

## id chol stab.glu hdl ratio glyhb location age gender height weight frame  
## 1 1000 203 82 56 3.6 4.31 Buckingham 46 female 62 121 medium  
## 2 1001 165 97 24 6.9 4.44 Buckingham 29 female 64 218 large  
## 3 1002 228 92 37 6.2 4.64 Buckingham 58 female 61 256 large  
## 4 1003 78 93 12 6.5 4.63 Buckingham 67 male 67 119 large  
## 5 1005 249 90 28 8.9 7.72 Buckingham 64 male 68 183 medium  
## 6 1008 248 94 69 3.6 4.81 Buckingham 34 male 71 190 large  
## bp.1s bp.1d bp.2s bp.2d waist hip time.ppn  
## 1 118 59 NA NA 29 38 720  
## 2 112 68 NA NA 46 48 360  
## 3 190 92 185 92 49 57 180  
## 4 110 50 NA NA 33 38 480  
## 5 138 80 NA NA 44 41 300  
## 6 132 86 NA NA 36 42 195

The goal of the analysis is to estimate a descriptive model to describe the variation of glyhb as a linear function of hdl, chol, age, gender, waist circumference, bp.1s (systolic blood pressure) and body-mass-index (weight in kg/height in m squared; ).

### Basic description of data set

1. What is the number of observations and what is the number of variables in that data set?

n\_observations <- nrow(diabetes)  
n\_variables <- ncol(diabetes)

Fill in:

* Number of observations: 403
* Number of variables: 19

### Generate new variable BMI

1. Generate a new variable BMI in the data set by computing it from weight and height. Note that height is given in inches and weight in lb. The correct formula is given by .

diabetes$BMI <- diabetes$weight / diabetes$height^2 \* 704.5

### Missing values

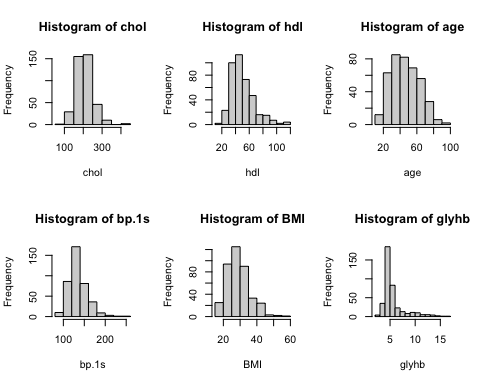
1. Investigate the number of missing values per variable. Compute the number of complete observations with respect to the variables chol, hdl, age, gender, bp1.s, BMI, and glyhb.

n\_missing\_entries <- sapply(diabetes, function(x) sum(is.na(x)))  
n\_complete\_obsersavtions <- n\_observations - n\_missing\_entries  
independent\_variables = c("chol", "hdl", "age", "gender", "bp.1s", "BMI", "glyhb")  
print(n\_complete\_obsersavtions[independent\_variables])

## chol hdl age gender bp.1s BMI glyhb   
## 402 402 403 403 398 397 390

1. Generate histograms for all independent variable of the above-described descriptive model. Generate a table with means, standard deviations and number and proportion missing for all variables, using the function of the RStudio class (mean.sd).

par(mfrow = c(2, 3))  
with(diabetes, {  
 hist(chol)  
 hist(hdl)  
 hist(age)  
 hist(bp.1s)  
 hist(BMI)  
 hist(glyhb)  
})



par(mfrow = c(1, 1))  
  
  
independent\_variable\_data <- diabetes[independent\_variables]  
independent\_variable\_data$gender[independent\_variable\_data$gender=="female"] <- 1  
independent\_variable\_data$gender[independent\_variable\_data$gender=="male"] <- 0  
independent\_variable\_data$gender <- as.numeric(independent\_variable\_data$gender)  
  
table = data.frame(  
 mean = sapply(independent\_variable\_data, function(x) mean(x, na.rm=T)),  
 sd = sapply(independent\_variable\_data, function(x) sd(x, na.rm=T)),  
 percentage\_missing = sapply(independent\_variable\_data, function(x) mean(100\*is.na(x)))  
)  
  
print(table)

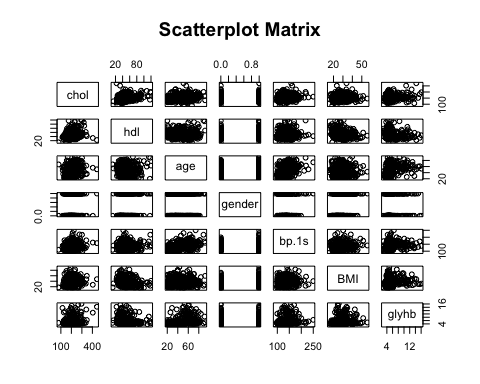
## mean sd percentage\_missing  
## chol 207.8457711 44.4455574 0.248139  
## hdl 50.4452736 17.2626257 0.248139  
## age 46.8511166 16.3123330 0.000000  
## gender 0.5806452 0.4940669 0.000000  
## bp.1s 136.9045226 22.7410332 1.240695  
## BMI 28.8459882 6.6199849 1.488834  
## glyhb 5.5897692 2.2425948 3.225806

1. Investigate the correlation between the independent variables by means of Spearman’s correlation coefficients. You can also use code similar to the lecture, where a matrix of scatterplots and correlation coefficients is produced.

cor(independent\_variable\_data, use="pairwise.complete.obs", method="spearman")

## chol hdl age gender bp.1s BMI  
## chol 1.0000000 0.14575720 0.26830889 0.01451690 0.20119153 0.11566877  
## hdl 0.1457572 1.00000000 -0.03716087 0.13734936 -0.04208816 -0.25169915  
## age 0.2683089 -0.03716087 1.00000000 -0.07793081 0.45457502 0.01342476  
## gender 0.0145169 0.13734936 -0.07793081 1.00000000 -0.06676174 0.25804507  
## bp.1s 0.2011915 -0.04208816 0.45457502 -0.06676174 1.00000000 0.15651689  
## BMI 0.1156688 -0.25169915 0.01342476 0.25804507 0.15651689 1.00000000  
## glyhb 0.2293256 -0.19241344 0.43041371 -0.06153921 0.28493979 0.21198669  
## glyhb  
## chol 0.22932560  
## hdl -0.19241344  
## age 0.43041371  
## gender -0.06153921  
## bp.1s 0.28493979  
## BMI 0.21198669  
## glyhb 1.00000000

pairs(formula = ~chol+hdl+age+gender+bp.1s+BMI+glyhb , data = independent\_variable\_data, main = "Scatterplot Matrix")



1. Provide an executive summary of the initial data analysis!

Your executive summary of the initial data analysis:

The univariate distributions of the continuous independent variables chol, age, BMI and bp.1s are approximately normally distributed according to visual inspection of the histograms in 4. The univariate distributions of hdl and glyhb look skewed.

Most variables have no or only weak correlations. However, stronger correlation was found between age and bp.s1/glyhb with Spearman correlation coefficients of 0.45/0.43 (see 5.).

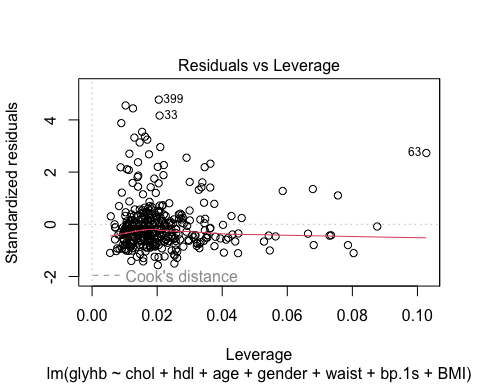
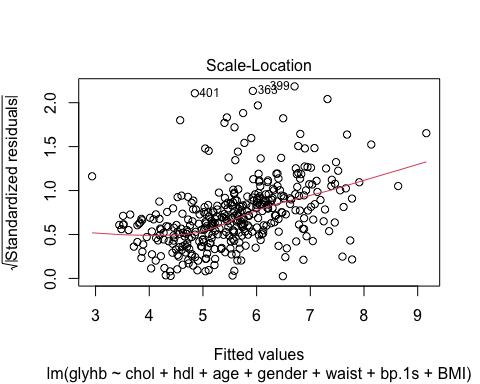
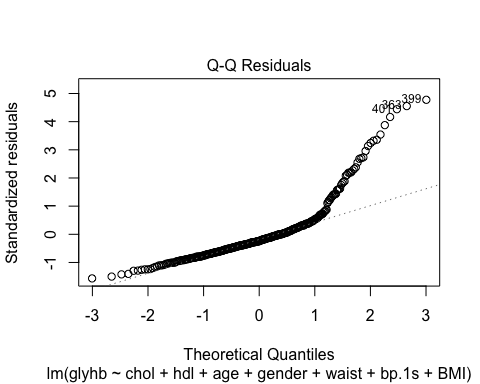
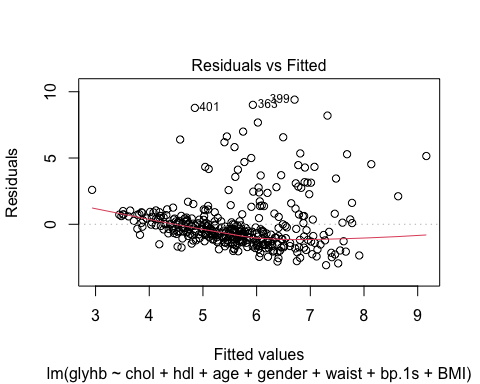
There are no missing values for age and gender and only very little missing values for the remaining independent variables chol, hdl, bp1.s and BMI as well as for the dependent variable glyhb (see 3.).

1. Estimate the linear regression model glyhb ~ chol + hdl + age + gender + waist + bp.1s + BMI. Provide a summary of the model. Investigate the residuals of the model. What are possible violations of this model’s assumptions?

model <- lm(formula = glyhb ~ chol + hdl + age + gender + waist + bp.1s + BMI, data=diabetes)  
summary(model)

##   
## Call:  
## lm(formula = glyhb ~ chol + hdl + age + gender + waist + bp.1s +   
## BMI, data = diabetes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0836 -1.1539 -0.4747 0.4386 9.4034   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.514291 1.003992 0.512 0.60879   
## chol 0.010906 0.002456 4.440 1.19e-05 \*\*\*  
## hdl -0.021068 0.006401 -3.292 0.00109 \*\*   
## age 0.031905 0.007228 4.414 1.34e-05 \*\*\*  
## gendermale -0.024528 0.228228 -0.107 0.91447   
## waist 0.084642 0.034069 2.484 0.01342 \*   
## bp.1s 0.002039 0.005065 0.403 0.68745   
## BMI -0.038170 0.029936 -1.275 0.20310   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.989 on 369 degrees of freedom  
## (26 observations deleted due to missingness)  
## Multiple R-squared: 0.2075, Adjusted R-squared: 0.1924   
## F-statistic: 13.8 on 7 and 369 DF, p-value: 6.864e-16

plot(model)



print(mean(residuals(model)))

## [1] -2.12032e-17

Your conclusions about possible violations of assumptions of the model:

The QQ-Plot shows that the residuals are not normally distributed. The scale-location plot shows that the residuals tend to get larger for larger fitted values, therefore it seems like the homoscedacity assumption is mildly violated. Thus, not all assumptions of the linear regression model (as defined on slide 35, lecture 1) are met.