Assignment 5

Luis Nicolas Luarte Rodriguez

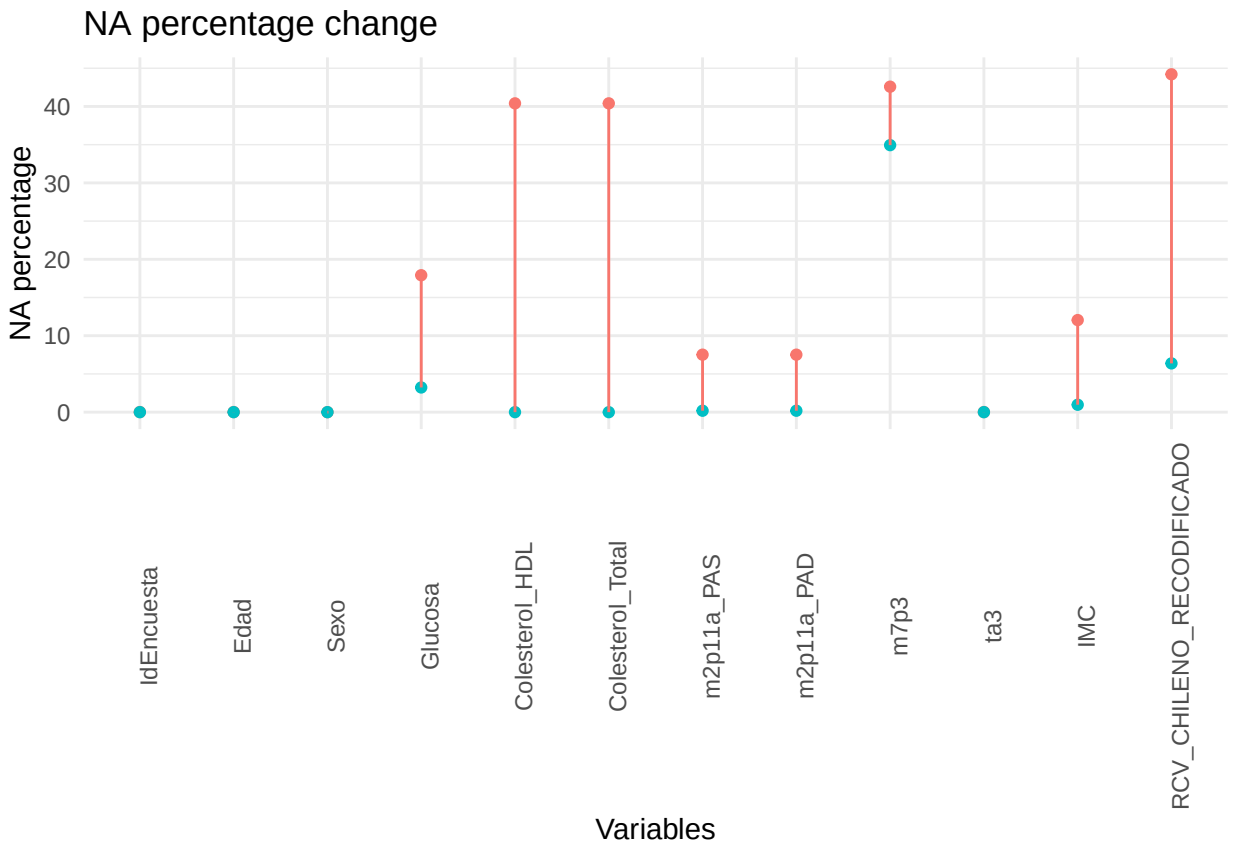
```
# load libs
if (!require("pacman")) install.packages("pacman")
pacman::p_load(survival, knitr, ggplot2, knitr, kableExtra, tidyverse,
    My.stepwise, reshape2, GGally, ggpubr, gtsummary, formatR, gt, olsrr,
    stargazer, boot)
```

```
# load the dataset
setwd("/home/nicoluarte/uni/PHD/stat_course")
dataSet <- as_tibble(read.csv("ENS.csv", sep = ",", header = TRUE))
t <- as.data.frame(read.csv("ENS.csv", sep = ",", header = TRUE))
head(dataSet)
```

1. Begin by examining the ENS 2016-17 database, measured variables, number of subjects examined, missing data

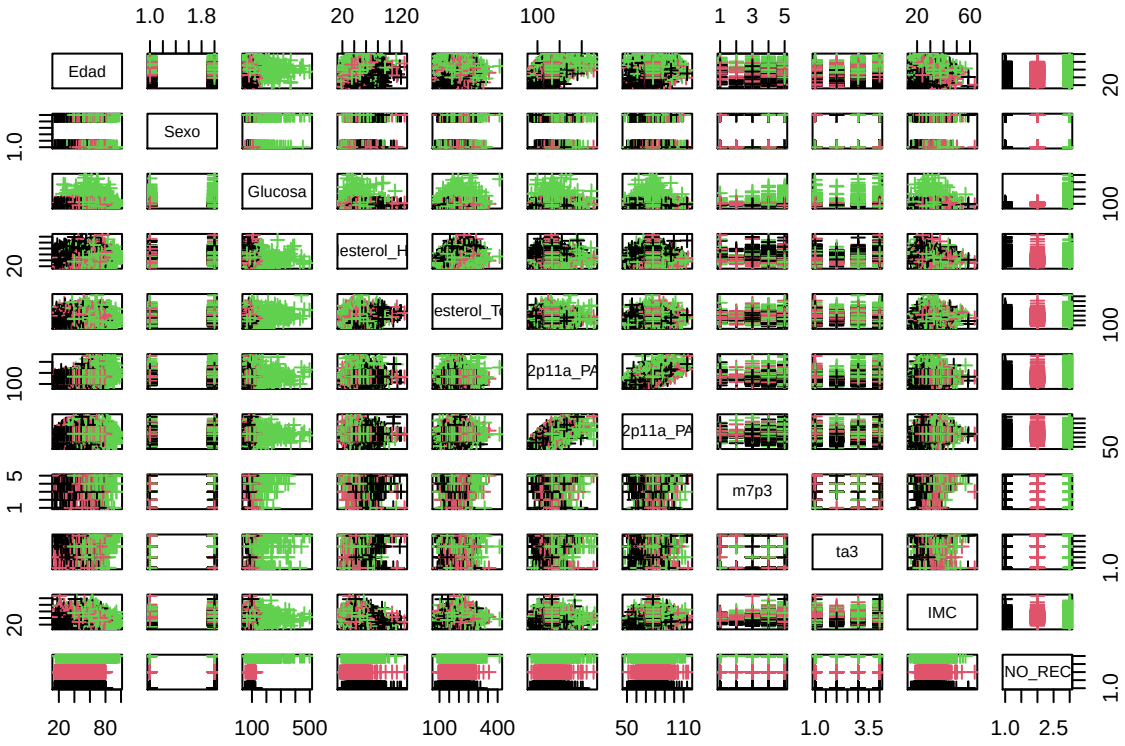
a. What is the percentage of missing data for each of the variables that appear in table 1. Remove all subjects for which HDL value is missing. What is now the percentage of missing data for each of the variables? From now on we will work with this subset of data.

```
# first I correct to the correct datatype
dataSetClean <- dataSet %>% mutate_at(vars(-"Sexo"), as.numeric) %>% mutate_at(vars("Sexo"),
    as.factor)

# then I get the percentage of NA
naPercentage <- dataSetClean %>% map(~mean(is.na(.)) * 100)

# remove NA values in HDL
dataSetFinal <- dataSetClean %>% drop_na(Colesterol_HDL)
# get percentage of missing data
naPercentageFinal <- dataSetFinal %>% map(~mean(is.na(.)) * 100)

initial <- melt(data.frame(naPercentage))
final <- melt(data.frame(naPercentageFinal))
plotDf <- melt(data.frame(x = initial$variable, y1 = initial$value, y2 = final$value))

# plot the differences

ggplot(plotDf, aes(x = x, y = value, colour = variable)) + geom_point() +
    geom_line(aes(group = x)) + labs(title = "NA percentage change") +
    xlab("Variables") + ylab("NA percentage") + theme_minimal() + theme(legend.position = "none") +
    theme(axis.text.x = element_text(angle = 90))
```


NA percentage change
NA percentage
40
30
20
10
0
IdEncuesta
Edad
Sexo
Glucosa
Colesterol_HDL
Colesterol_Total
m2p11a_PAS
m2p11a_PAD
m7p3
ta3
IMC
RCV_CHILENO_RECODIFICADO
Variables

b. Plot matrix of scatterplots for the value of each of the 10 relevant variables similar to the figure 2, assigning a different color to each of the subjects depending on their class regarding the RCV (recommendation: use the command gplotmatrix). Include the name of the variables in the diagonal and increase the size of the point if the graph with the default point size cannot be apreciated.

```
pairs(dataSetFinal[, 2:12], col = as.factor(dataSetFinal$RCV_CHILENO_RECODIFICADO),
    pch = 3)
```



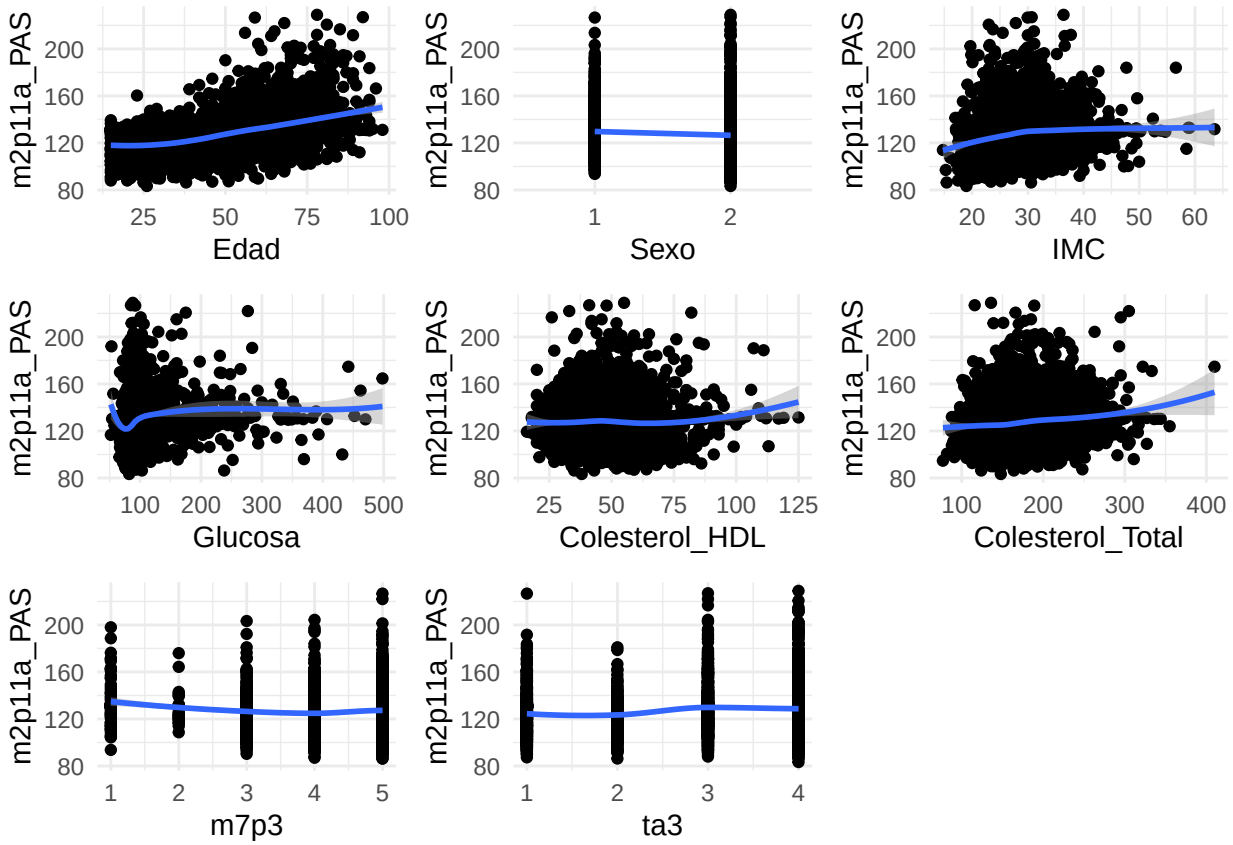
c. Interpret the results of the scatterplots. Can you identify at first glance the variables with greater separation between the different categories? Can you identify correlations between some of the variables?

At first glance glucose seems to separate most of the variables, higher levels of glucose are associated with cardiovascular risk '3' (likely, higher risk of cardiovascular failure). Thus, glucose is associated with higher cardiovascular risk. Age also separate data quite clearly, older people score, more frequently, '3' in cardiovascular risk. As for correlations, systolic and diastolic blood pressure show the highest correlation; both type of cholesterol show a similar trend; age and sys/diastolic blood pressure show a positive correlation, but weaker than the previously mentioned. Regarding negative correlations, a weak one is the present between HDL cholesterol and IMC.

2. Build a linear regression model to predict the systolic blood pressure

a. Use a multiple linear regression model with age, sex, BMI, glucose, total cholesterol, HDL cholesterol, alcohol and tobaco consumption as predictors and present the results including the confidence interval for the parameters and the value of R^2 (variance explained by the model). Explain how you are handling missing data.

```
# check multiple linear regression assumptions linear relationship
# between dependent and independent variables
dataModel <- dataSetFinal %>% drop_na() %>% select(m2p11a_PAS, Edad, Sexo,
    IMC, Glucosa, Colesterol_HDL, Colesterol_Total, m7p3, ta3)

# plot against every predictor
dVars <- colnames(dataModel)[-1]
pData <- function(x) {
    ggplot(dataSetFinal, aes(x = .data[[x]], y = m2p11a_PAS)) + geom_point() +
        geom_smooth(method = "loess", aes(group = "1")) + theme_minimal()
}
plots <- dVars %>% map(~pData(.x))
ggarrange(plotlist = plots)
```




```
# relationships are somewhat linear, no clearly curvilinear
# relationship detected

# fit the model
mdl <- lm(data = dataModel, m2p11a_PAS ~ .)
summary(mdl)
```

```
fit_stats <- broom::glance(mdl) %>% select(`R<sup>2</sup>` = r.squared,
    AIC) %>% mutate_all(function(x) style_sigfig(x, digits = 3)) %>% {
    paste(names(.), ., sep = " = ", collapse = "; ")
}
# generate table with results
stargazer(mdl, header = FALSE, type = "latex", title = "Regression Results")
```

All rows containing 1 or more missing values were eliminated to fit the model.

b. Interpret the results obtained: what are the variables that have significant contribution to the model? Was this result expected?

In general the model does not fit data very well, r squared measure of fit is below 0.3. Thus, only about 30% of the variance is explained by the model. Only age, sex, BMI, HDL, and total cholesterol reach statistical significance. Among significant covariates, the model predicts that systolic blood pressure increase as one gets older ($\beta = 0.37$, $p =< 0.001$). The same relationship is followed by BMI ($\beta = 0.43$, $p =< 0.001$), and to a lesser extent by HDL and total cholesterol $\beta = 0.06$ and $\beta = 0.02$, respectively. As for sex, code '2' (possibly female), shows a reduction in systolic blood pressure compared with code '1' ($\beta = -6.4$, $p =< 0.001$), this is the strongest predictor for systolic pressure. I was expecting a better fit, mostly regarding cholesterol as is typically associated with blood pressure problems. Perhaps this relationship is heavily attenuated when controlling for all the other variables.

Table 1: Regression Results

	Dependent variable:
	m2p11a_PAS
Edad	0.373^{***}
	(0.017)
Sexo2	-6.382^{***}
	(0.635)
IMC	0.427^{***}
	(0.060)
Glucosa	0.017^{*}
	(0.010)
Colesterol_HDL	0.063^{***}
	(0.025)
Colesterol_Total	0.022^{***}
	(0.008)
m7p3	0.191
	(0.329)
ta3	0.306
	(0.247)
Constant	89.893^{***}
	(2.905)
Observations	2,251
R^2	0.270
Adjusted R^2	0.267
Residual Std. Error	13.977 (df = 2242)
F Statistic	103.556^{***} (df = 8; 2242)
Note:	*p<0.1; **p<0.05; ***p<0.01

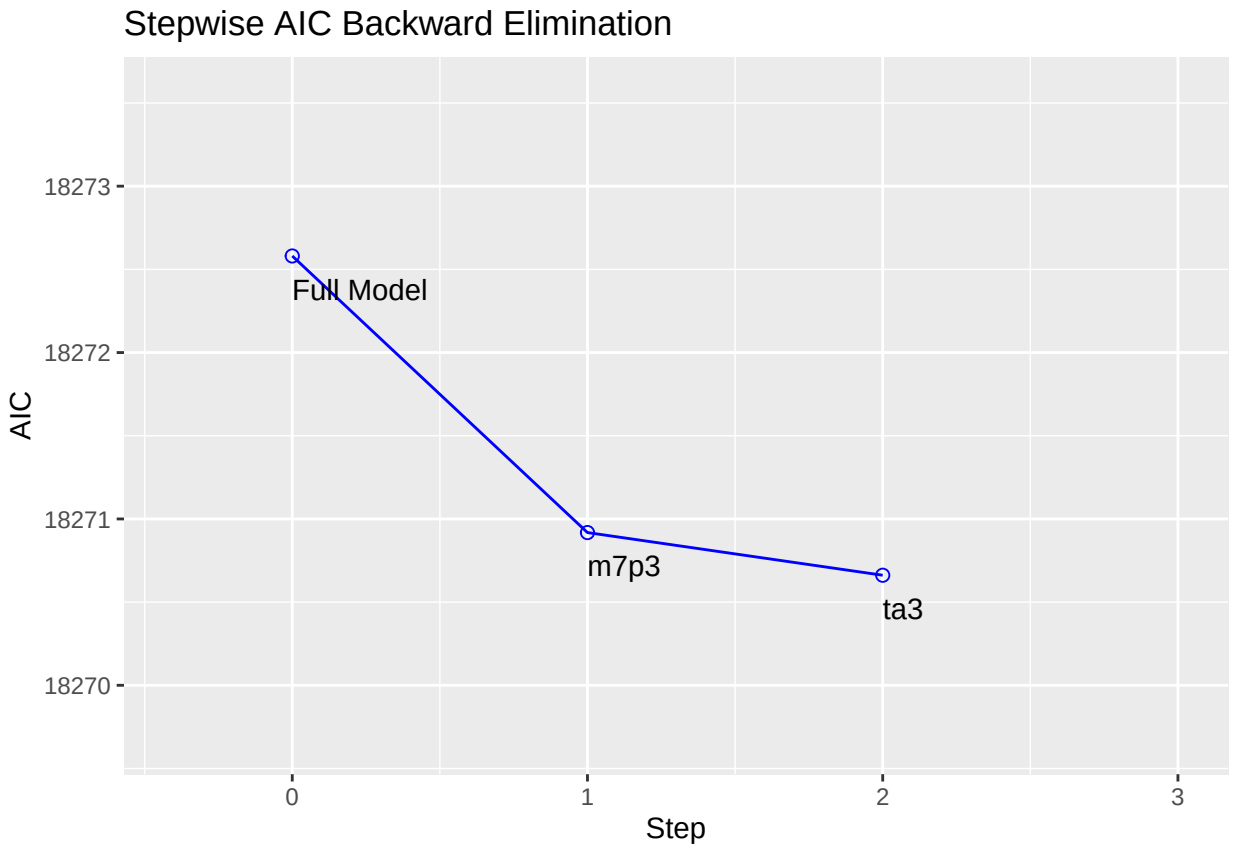
c. How does a linear regression model predict the systolic blood pressure? Comment the results.

As previously stated the model does not provide a good fit. Sex was one of the most significant covariates, which make sense because between the variables included it refers directly to genetic differences. BMI is somewhat controlled between males and females because it considers height, which is, typically, a variable that is quite different between males and females. The other variables are likely to be related to life-style choices, which should not be so different between males and females. Finally, the low r squared might be related to systolic blood pressure not being a very accurate measure, because of daily fluctuations.

3. In the previous section we have seen that there are some variables that do not have a significant contribution to the model:

a. Select the best model using "backward selection"

```
# backwards assuming the previous model is the full model
finalModel <- ols_step_backward_aic(mdl, details = TRUE)
plot(finalModel)
```




```
summary(finalModel$model)
```

b. Compute the confidence intervals for the parameters of the best model using the Wald method seen in class (coefCI(mdl) in Matlab)

```
# confidence interval for the parameter using Walds
model_ci <- confint(finalModel$model, level = 0.95, method = c("Wald"))
stargazer(model_ci, header = FALSE, type = "latex", title = "Wald CI")
```

Table 2: Wald CI

	2.5 %	97.5 %
(Intercept)	86.302	96.266
Edad	0.343	0.410
Sexo2	-7.477	-5.056
IMC	0.315	0.549
Glucosa	-0.003	0.038
Colesterol_HDL	0.017	0.112
Colesterol_Total	0.006	0.037

c. Compute again the confidence intervals using bootstrap and compare the results

```
bootFunc <- function(data, idx) {
    coef(lm(m2p11a_PAS ~ Edad + Sexo + IMC + Glucosa + Colesterol_HDL +
        Colesterol_Total, data = dataModel[idx, ]))
}
bootRes <- boot(data = dataModel, statistic = bootFunc, R = 1000)
bootCI <- 1:length(coef(finalModel$model)) %>% map(function(x) as.vector(boot.ci(bootRes,
    type = "perc", index = x)$perc)[-3:-1]) %>% unlist() %>% matrix(.,
    nrow = length(.)/2, byrow = T) %>% as.data.frame(.) %>% mutate(Coefficients = names(coef(finalModel$
    select(Coefficients, everything()) %>% rename(., `2.5%` = V1, `97.5%` = V2)
knitr::kable(bootCI, caption = "Bootstrap CI") %>% kable_styling(latex_options = "hold_position")
```

Table 3: Bootstrap CI

Coefficients	2.5%	97.5%
(Intercept)	86.2298977	96.4394903
Edad	0.3374355	0.4127881
Sexo2	-7.5431388	-5.0025631
IMC	0.3118954	0.5525700
Glucosa	-0.0068741	0.0458046
Colesterol_HDL	0.0153008	0.1128265
Colesterol_Total	0.0051792	0.0375856

Coefficients confidence intervals are almost the same in all cases. Coefficients values were obtained from 1000 bootstrap replicated and converged into similar values. The method used for obtaining the confidence intervals was the percentile bootstrap.