Modelos Lineares PPGEPI

Gabriela Wünsch Lopes e Danilo de Paula Santos

02/12/2019

#Instruções:

Cada dupla deverá escolher uma base de dados, que pode ser analisado via modelos lineares (um desfecho quantitativo e no mínimo dois preditores).

Proceder a análise descritiva, modelagem e análise de diagnóstico. Preparar o parágrafo da análise estatística conforme seria necessário para a seção de Métodos para um artigo.

Construir as tabelas e/ou figuras, e escrever as interpretações, como para uma sessão de Resultados de um artigo.

Preparar uma apresentação de 15 minutos para dia 06/12.

Postar o arquivo em Word, a apresentação e os dados em SPSS/XLS/CSV. Referir a fonte dos dados.

# install.packages("ggplot2")  
# install.packages("car")  
# install.packages("tidyverse")  
# install.packages("mctest")  
  
  
library("ggplot2")  
library("car")

## Loading required package: carData

library("tidyverse")

## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --

## v tibble 2.1.3 v purrr 0.3.2  
## v tidyr 0.8.3 v dplyr 0.8.3  
## v readr 1.3.1 v stringr 1.4.0  
## v tibble 2.1.3 v forcats 0.4.0

## -- Conflicts ------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::recode() masks car::recode()  
## x purrr::some() masks car::some()

library("mctest")  
library("compareGroups")

## Loading required package: SNPassoc

## Loading required package: haplo.stats

## Loading required package: survival

## Loading required package: mvtnorm

## Loading required package: parallel

## Registered S3 method overwritten by 'SNPassoc':  
## method from   
## summary.haplo.glm haplo.stats

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

# Descrição do banco:

## Diabetes Dataset:

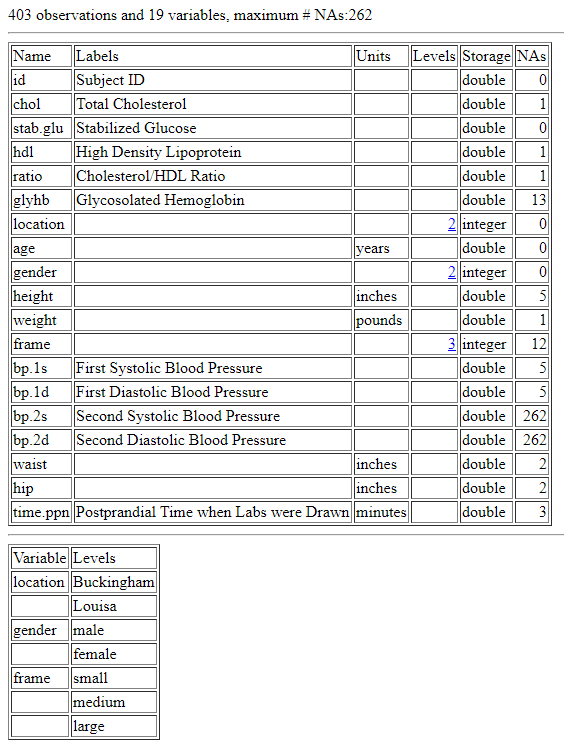
These data are courtesy of Dr John Schorling, Department of Medicine, University of Virginia School of Medicine. The data consist of 19 variables on 403 subjects from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. According to Dr John Hong, Diabetes Mellitus Type II (adult onset diabetes) is associated most strongly with obesity. The waist/hip ratio may be a predictor in diabetes and heart disease. DM II is also agssociated with hypertension - they may both be part of “Syndrome X”. The 403 subjects were the ones who were actually screened for diabetes. Glycosolated hemoglobin > 7.0 is usually taken as a positive diagnosis of diabetes. For more information about this study see

Willems JP, Saunders JT, DE Hunt, JB Schorling: Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. Southern Medical Journal 90:814-820; 1997

and

Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL: A trial of church-based smoking cessation interventions for rural African Americans. Preventive Medicine 26:92-101; 1997.

## Documentação banco:



“Documentação banco”

dados\_dm <- read.csv("diabetes.csv")  
  
# convertendo as medidas do sistema imperial para o métrico:  
  
dados\_dm$peso\_kg <- dados\_dm$weight \* 0.453592  
  
dados\_dm$altura\_cm <- dados\_dm$height \* 2.54  
  
dados\_dm$cintura\_cm <- dados\_dm$waist \* 2.54  
  
dados\_dm$quadril\_cm <- dados\_dm$hip \* 2.54  
  
dados\_dm$taxa\_hdl <- 1/dados\_dm$ratio  
  
dados\_dm$imc <- dados\_dm$peso\_kg / I((dados\_dm$altura\_cm/100) ^2)

Criando uma variável de status de diabetes:

0 = Não diabético 1 = Diabético

dados\_dm$diabetes <- factor(ifelse(is.na(dados\_dm$glyhb), NA,  
 ifelse(dados\_dm$glyhb >= 6.5, 1, 0)))

# Hipótese:

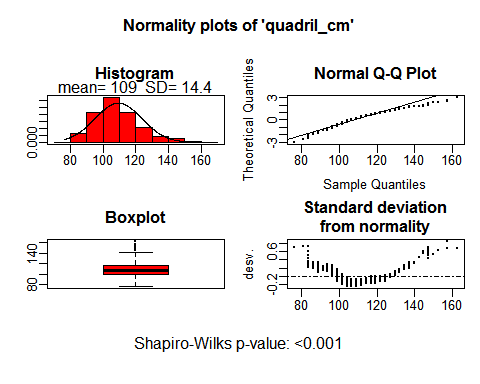
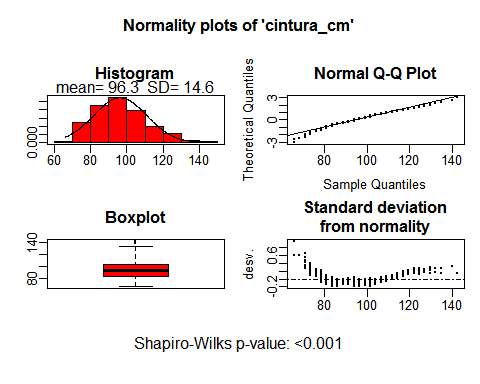
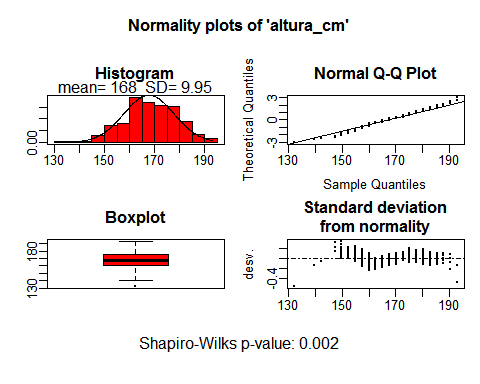
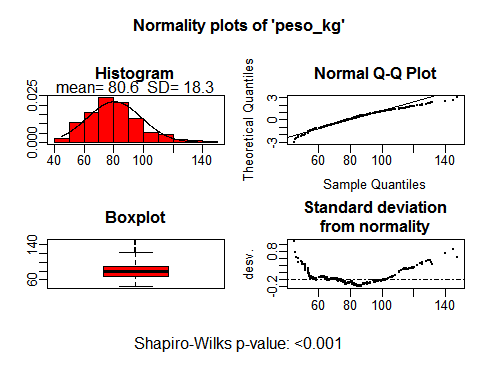
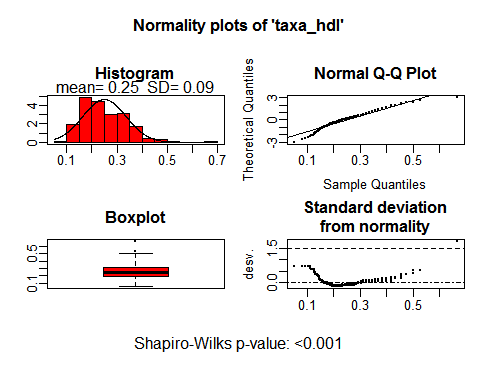
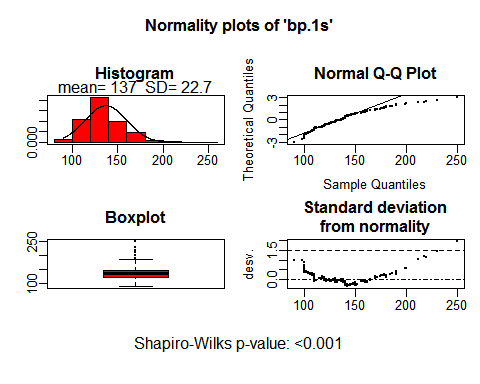
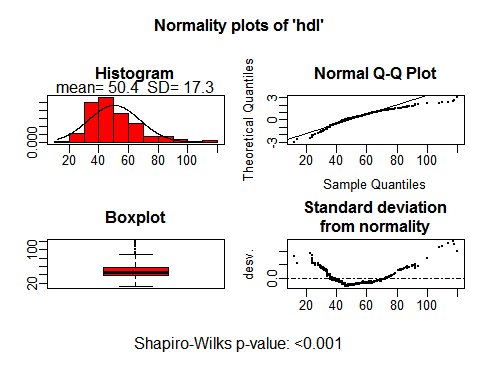
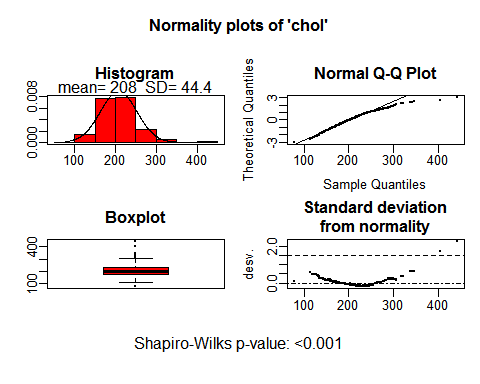
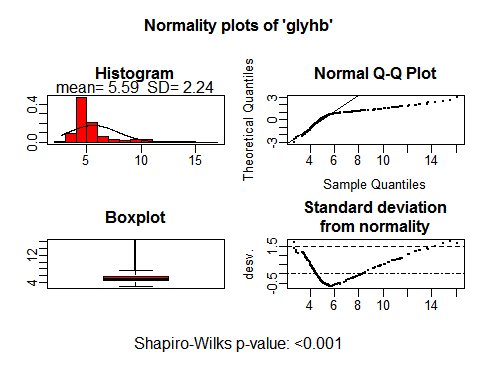
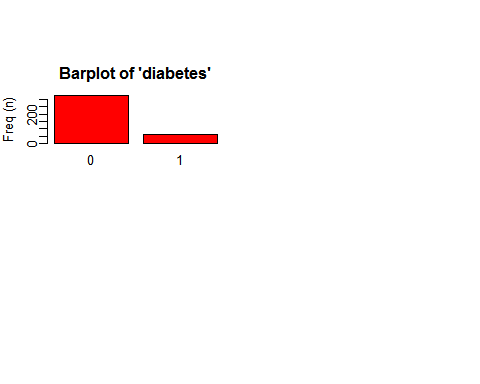
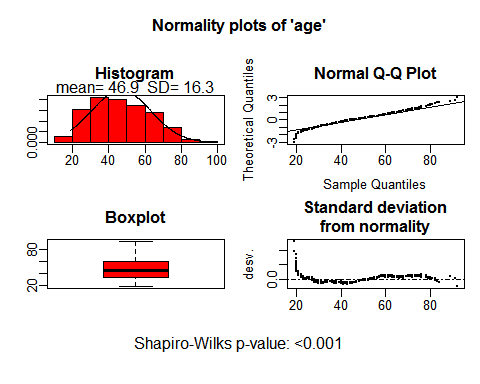
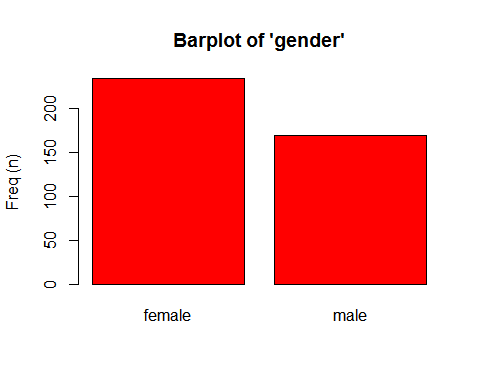
Níveis séricos mais altos de colesterol HDL se associam com um melhor controle glicêmico.

compare\_dm <- compareGroups(data = dados\_dm, ~ gender + age + diabetes + glyhb + chol + hdl + bp.1s + + taxa\_hdl + peso\_kg + altura\_cm + cintura\_cm + quadril\_cm )  
  
tabela\_dm <- createTable(compare\_dm )  
  
export2md(tabela\_dm)

Summary descriptives table

|  |  |  |
| --- | --- | --- |
|  | [ALL] N=403 | N |
| gender: |  | 403 |
| female | 234 (58.1%) |  |
| male | 169 (41.9%) |  |
| age | 46.9 (16.3) | 403 |
| diabetes: |  | 390 |
| 0 | 325 (83.3%) |  |
| 1 | 65 (16.7%) |  |
| glyhb | 5.59 (2.24) | 390 |
| chol | 208 (44.4) | 402 |
| hdl | 50.4 (17.3) | 402 |
| bp.1s | 137 (22.7) | 398 |
| taxa\_hdl | 0.25 (0.09) | 402 |
| peso\_kg | 80.6 (18.3) | 402 |
| altura\_cm | 168 (9.95) | 398 |
| cintura\_cm | 96.3 (14.6) | 401 |
| quadril\_cm | 109 (14.4) | 401 |

plot(tabela\_dm)

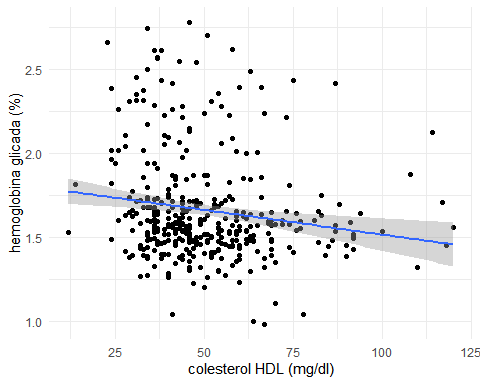


Avaliação da linearidade da relação entre colesterol HDL e hemoglobina glicada:

scatter\_hdl <- ggplot(dados\_dm, aes(x = hdl, y = log(glyhb))) +  
 geom\_point() +  
 scale\_x\_continuous("colesterol HDL (mg/dl)") +  
 scale\_y\_continuous("hemoglobina glicada (%)") +  
 theme\_minimal() +  
 geom\_smooth(method = "lm")  
scatter\_hdl

## Warning: Removed 14 rows containing non-finite values (stat\_smooth).

## Warning: Removed 14 rows containing missing values (geom\_point).



# Modelo inicial, apenas com HDL:

lm\_hdl <- lm(formula = (glyhb) ~ hdl,  
 data = dados\_dm)  
  
summary(lm\_hdl)

##   
## Call:  
## lm(formula = (glyhb) ~ hdl, data = dados\_dm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9224 -1.2092 -0.7061 0.0686 10.4343   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.565734 0.346607 18.943 < 2e-16 \*\*\*  
## hdl -0.019348 0.006521 -2.967 0.00319 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.222 on 387 degrees of freedom  
## (14 observations deleted due to missingness)  
## Multiple R-squared: 0.02224, Adjusted R-squared: 0.01972   
## F-statistic: 8.804 on 1 and 387 DF, p-value: 0.003192

## Montando um novo modelo, agora com ajuste para outras variáveis

* Nosso método de modelagem será o stepwise baseado R2 ajustado.

Incluindo idade e sexo, variáveis que serão mantidas no modelo independentemente de significância estatística:

lm\_hdl1 <- lm(formula = glyhb ~ hdl + age + gender,  
 data = dados\_dm)  
  
summary(lm\_hdl1)

##   
## Call:  
## lm(formula = glyhb ~ hdl + age + gender, data = dados\_dm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1135 -1.1807 -0.4631 0.2944 10.1207   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.423480 0.448907 9.854 < 2e-16 \*\*\*  
## hdl -0.020514 0.006171 -3.324 0.000972 \*\*\*  
## age 0.046796 0.006472 7.230 2.62e-12 \*\*\*  
## gendermale 0.029318 0.217179 0.135 0.892686   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.089 on 385 degrees of freedom  
## (14 observations deleted due to missingness)  
## Multiple R-squared: 0.1403, Adjusted R-squared: 0.1336   
## F-statistic: 20.94 on 3 and 385 DF, p-value: 1.372e-12