Análise descritiva e estatística de uma base de dados de salários

Otto Tavares

28 March, 2024

## Introdução - Bibliotecas (parte 1)

Carregando bibliotecas que foram fundamentais para a construção dos modelos de regressão, tanto na versão com *input* de dados faltantes, como nos modelos usuais.

library(tidyverse)  
library(tidyr)  
library(purrr)  
library(dlookr)  
library(summarytools)  
library(readxl)  
library(knitr)  
library(data.table)  
library(ggpubr)  
library(corrplot)

## Introdução - Bibliotecas (parte 2)

library(rcompanion)  
library(stargazer)  
library(mice)  
library(rmarkdown)  
library(tinytex)  
library(sandwich)  
library(magrittr)  
library(shiny)  
library(plm)

# Base de dados

## Base de dados de trabalho

* Base de dados do curso está disponibilizada no github no diretório ‘dados\_auxiliares’.
* As bases disponíveis até aqui são:
  1. As de população mundial extraída do wikipedia;
  2. Lista dos países por continente no mundo;
  3. Salários extraídas do Livro do Bussab e Moretim;
  4. Crimes extraída do Instituto de Segurança Pública;
  5. Income, disponbilizada pelos autores Acemoglu e Robinson.
* Vamos importar a base de salários para exposição das estatísticas descritivas em relatório em slides.

## Imprindo as duas primeiras linhas da base de salário

* Apresentando as primeiras linhas do banco de dados de Salários para termos ciência dos dados.

## Identificando os tipos de cada variável na base

* A função diagnose que utilizamos para identificar tipos de variável, sua unicidade e proporção de *missing*.

salarios %>% dlookr::diagnose() %>%   
 kable(., , format = "latex", booktabs = T) %>%  
 kableExtra::kable\_styling(font\_size = 7)

## Análise de frequências da variável com dados faltantes **n\_filhos**

* Utilizamos a função freq() do pacote summarytools para calcular as frequências relativas

## Análise descritiva e de histogramas de uma variável contínua

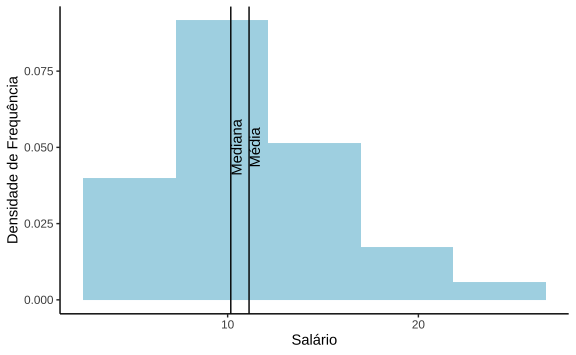
* Variável salários é analisada descritivamente.
* A centralidade dos dados, a dipersão, a assimetria, bem como as estatísticas de ordem são calculadas, a fim de ter uma leitura acerca da distribuição dessa variável.

## Função de Sturge para cálculo do número de bins

sr <- function(x) {  
 n <-length(x)  
 return((3.49\*sd(x))/n^(1/3))  
}

## Análise visual da variável salário

* Calculando o histograma da variável salários com o número de *bins* calculado a partir da função de Sturge.



## Análise visual da variável salário - leitura

1. Leve assimetria com cauda à direita
2. Centralidade dos dados calculada pela média sofre leve contaminação dos valores mais distantes do centro da distribuição
3. Por mais que sejam poucas observações os dados não apresentam dispersão elevada, tendo a maioria dos dados concentrada próxima ao centro da distribuição.

## Rodando a regressao linear sem a variável n\_filhos

## Multivariada, com a variável estado civil de controle

## Multivariada, com as variáveis estado civil, grau de instrucao de controle

# Resultados das Regressões

Salário   
 Modelo 1 Modelo 2 Modelo 3

| Idade (anos) 0.247\*\* (0.109) 0.233\*\* (0.108) 0.345\*\*\* (0.071) Estado Civil (Solteiro) -1.955 (1.443) -1.144 (0.951) Grau de Instrução (Médio) 4.603\*\*\* (1.081) Grau de Instrução (Superior) 9.779\*\*\* (1.391) Constant 2.566 (3.831) 3.917 (3.914) -4.225 (2.886) |
| --- |
| Observations 36 36 36 R2 0.132 0.178 0.687 Adjusted R2 0.107 0.128 0.647 Residual Std. Error 4.336 (df = 34) 4.284 (df = 33) 2.726 (df = 31) F Statistic 5.172\*\* (df = 1; 34) 3.567\*\* (df = 2; 33) 17.024\*\*\* (df = 4; 31) ============================================================================================= Note: *p<0.1;* ***p<0.05;*** p<0.01 |
| Resultados das Regressões === 1pt |

## Multivariada, com as variáveis estado civil, grau de instrucao e regiao de controle