Análise descritiva e estatística de uma base de dados de salários

Otto Tavares

01 Apr, 2025

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)
```

Section 1

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.

n	estado_civil	Grau_de_instrucao	n_filhos	salario	idade_anos	idade_meses	regiao
1	solteiro	ensino fundamental	NA	4.00	26	3	interior
2	casado	ensino fundamental	1	4.56	32	10	capital
3	casado	ensino fundamental	2	5.25	36	5	capital
4	solteiro	ensino médio	NA	5.73	20	10	outra
5	solteiro	ensino fundamental	NA	6.26	40	7	outra
6	casado	ensino fundamental	0	6.66	28	0	interior

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)
```

variables	types	missing_count	missing_percent	unique_count	unique_rate
n	numeric	0	0.00000	36	1.0000000
estado_civil	character	0	0.00000	2	0.0555556
Grau_de_instrucao	character	0	0.00000	3	0.0833333
n_filhos	numeric	16	44.44444	6	0.1666667
salario	numeric	0	0.00000	36	1.0000000
idade_anos	numeric	0	0.00000	24	0.6666667
idade_meses	numeric	0	0.00000	12	0.3333333
regiao	character	0	0.00000	3	0.0833333

Análise de frequências da variável com dados faltantes
n_filhos

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

```
## Error in table(names(candidates))[["tested"]]: subscript ou
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
0	4	20	20	11.111111	11.11111
1	5	25	45	13.888889	25.00000
2	7	35	80	19.444444	44.44444
3	3	15	95	8.333333	52.77778
5	1	5	100	2.777778	55.55556
<NA>	16	NA	NA	44.444444	100.00000
Total	36	100	100	100.000000	100.00000

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.

```
## Error in table(names(candidates))[["tested"]]: subscript ou
```

Descriptive Statistics

salario

N: 36

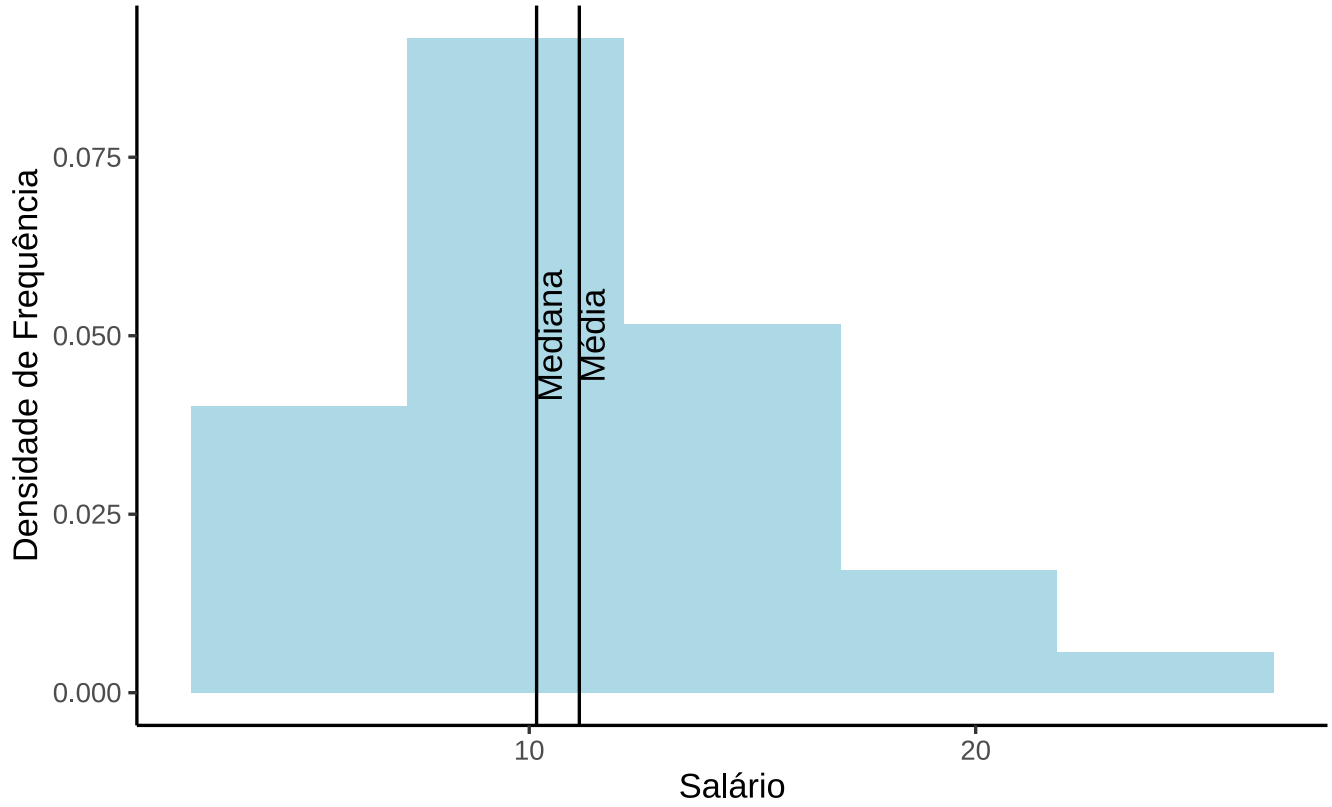
	salario
Mean	11.12
Std.Dev	4.59

Função de Sturge para cálculo do número de bins

```r
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

Resultados das Regressões

Salário

Modelo 1

Idade (anos)

0.247** (0.109)

Constant

2.566 (3.831)

R2

0.132

F Statistic

5.172** (df = 1; 34)

Note:

*p<0.1; **p<0.05; p<0.01

Multivariada, com a variável estado civil de controle

Table 2: Resultados das Regressões

	Salário	
	Modelo 1	Modelo 2
Idade (anos)	0.247** (0.109)	0.233** (0.108)
Estado Civil		−1.955 (1.443)
Constant	2.566 (3.831)	3.917 (3.914)
Observations	36	36
R^2	0.132	0.178
Adjusted R^2	0.107	0.128
Residual Std. Error	4.336 (df = 34)	4.284 (df = 33)
F Statistic	5.172** (df = 1; 34)	3.567** (df = 2; 33)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3: Resultados das Regressões

1pt

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

Table 4: 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
R^2	0.132	0.178	0.687
Adjusted R^2	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		

Table 5: Resultados das Regressões

1pt

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

Table 6: Resultados das Regressões

	Salário			
	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Idade (anos)	0.247** (0.109)	0.233** (0.108)	0.345*** (0.071)	0.351*** (0.074)
Estado Civil		−1.955 (1.443)	−1.144 (0.951)	−1.052 (1.010)
Grau de Instrução			4.603*** (1.081)	4.563*** (1.113)
Região			9.779*** (1.391)	9.757*** (1.431)
factor(regiao)interior				0.587 (1.190)
factor(regiao)outra				−0.019 (1.178)
Constant	2.566 (3.831)	3.917 (3.914)	−4.225 (2.886)	−4.638 (3.130)
Observations	36	36	36	36
R^2	0.132	0.178	0.687	0.691
Adjusted R^2	0.107	0.128	0.647	0.627
Residual Std. Error	4.336 (df = 34)	4.284 (df = 33)	2.726 (df = 31)	2.802 (df = 29)
F Statistic	5.172** (df = 1; 34)	3.567** (df = 2; 33)	17.024*** (df = 4; 31)	10.800*** (df = 6; 29)
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 7: Resultados das Regressões

1pt