DUAS VARIÁVEIS NUMÉRICAS

[revisar también este tutorial. Ver también Módulo Testes -- OLS:
http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r#kendall-rank-correlation-test
)

Nesta secção estudamos a relação entre duas variáveis quantitativas:

N -> N

Ferramentas principais que serão utilizadas:
- gráficos de dispersão
- coeficientes de correlação
- coeficientes de regressão

1. Antes da Análise: pergunta e dados

Pergunta de investigação

> *O nível del lealdade ao chefe do governo afecta a sua duraçaõ em funções dos ministros?*

A pergunta inclui uma variável independente numérica (nível de lealdade) e uma variável dependente numérica (número de dias no cargo de ministro).

Dados
Continuamos a trabalhar com os dados da base "BaseTutorial_1", agora focando-nos nas variáveis "lealdade" e "duração".

Controlamos a classe:

```
class(B$lealdade)
```

```
[1]"numeric"
```

Controlamos a escala da variável:

```
table (B$lealdade)
```

```
   0   0.25  0.5  0.75    1
   2    4    2     2      6
```

A variável regista 5 valores, numa escala de 0 a 1.

Controlamos pela existência de valores em falta (*missings*):

```
table (is.na(B$lealdade))

```

```
FALSE
   16
```

Olhamos para as suas estatísticas básicas:

```
summary(B$lealdade)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2500  0.6250  0.5938  1.0000  1.0000
```

2. Dados 'crus' (desagregados)

Os dados aparecem por indivíduo. O indivíduo 1 tem o nível máximo de lealdade e uma duração em funções de 200 dias, enquanto o indivíduo 16 tem um nível de lealdade de 0,50 e uma duração em funções de 110 dias.

Description: df [16 × 3]

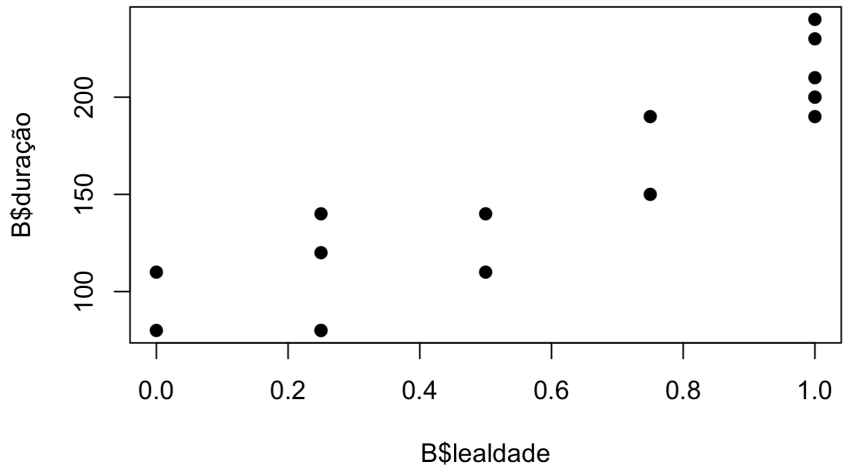
id <int>	lealdade <dbl>	duração <dbl>
1	1.00	200
2	1.00	240
3	0.50	140
4	0.25	120
5	0.75	150
6	1.00	230
7	1.00	190
8	0.25	140
9	0.00	80
10	1.00	210
11	1.00	200
12	0.25	80
13	0.25	80
14	0.00	110
15	0.75	190
16	0.50	110

16 rows

3. Análise

Gráfico de dispersão (scatterplot)

De acordo com a nossa pergunta de investigação, estamos interessados na interseção dos valores das duas variáveis para cada indivíduo. O gráfico de dispersão é a ferramenta básica para visualizar essa associação.



Tipicamente, a variável independente encontra-se no eixo horizontal X e a dependente no eixo vertical Y. Cada ponto representa um indivíduo, sendo que para cada indivíduo temos duas observações. Assim, cada ponto retrata a interSecção entre os valores obtidos por cada indivíduo nas duas variáveis.

Por exemplo, as duas primeiras observações do lado esquerdo mostram um nível de lealdade igual a 0 e uma duração em funções por volta dos 80 e dos 110 dias. Estas observações correspondem aos indivíduos 9 e 14 da base de dados. A diferença dos gráficos univariados, onde as observações são representadas segundo a sua localização na base de dados, é que neste caso as observações são ordenadas segundo as escalas das variáveis.

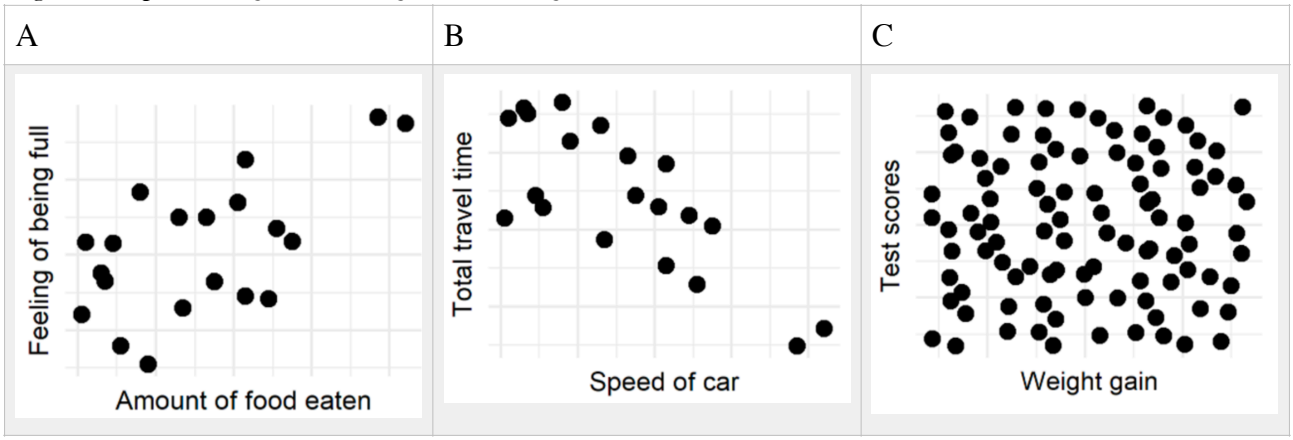
O gráfico de dispersão de pontos dá-nos informação sobre:
a) a direção da relação, que pode ser positiva, negativa, ou nula;
b) a forma da relação, que pode ser curvilínea, linear, ou cluster;
c) a força da associação entre as duas variáveis;
d) a presença de outliers que se "escapam" ao padrão global.

Direção, Forma e Força da associação

Se os valores da variável 1 aumentam quando os valores da variável 2 aumentam, estamos perante uma associação com **direção positiva**. Em termos gráficos, isto significa que os pontos no gráfico sobem no eixo do y à medida que sobem no eixo do x. Por exemplo, na figura A, a saciedade aumenta com a quantidade de alimentos ingeridos.
Se os valores da variável 1 diminuem à medida que os valores da variável 2 aumentam, isto é os pontos no gráfico vão descendo no eixo do y à medida que sobem no eixo do x do gráfico, estamos perante uma **associação negativa**. Na figura B, o tempo de viagem diminui com a velocidade do carro. Quanto mais veloz é o veículo menor é a duração da viagem.

Por fim, quando os pontos do gráfico ocupam posições altas e baixas ao longo de todo o eixo do x, possivelmente estão perante um caso em que as variáveis são independentes (ou talvez seja necessário controlar o efeito de outras variáveis para visualizar a associação). Na figura C, as notas no exame não variam em função do peso.

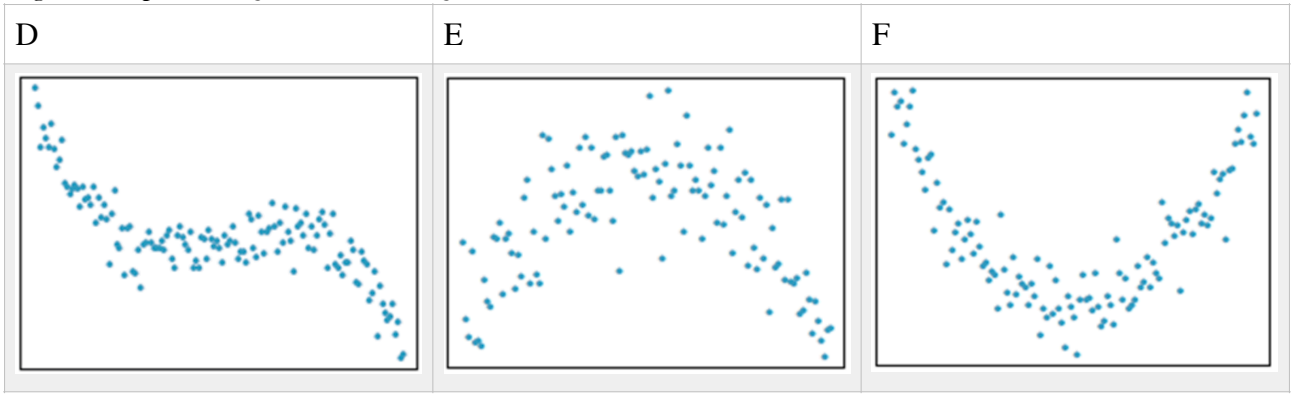
Figura: Representação de direções da relação



Fonte: https://www.freecodecamp.org/news/what-is-a-correlation-coefficient-r-value-in-statistics-explains/

Outro aspeto que se pode apurar é a **forma** da associação. Se a dispersão dos pontos descreve, em maior ou menor grau, uma linha reta a relação entre as duas variáveis diz-se linear. Caso contrário, diz-se não linear (figura D). Se a dispersão dos pontos representa um padrão em U, em posição regular ou invertida, a relação diz-se curvilínea (figuras E e F). De notar que estes dois casos representam igualmente relações não lineares.

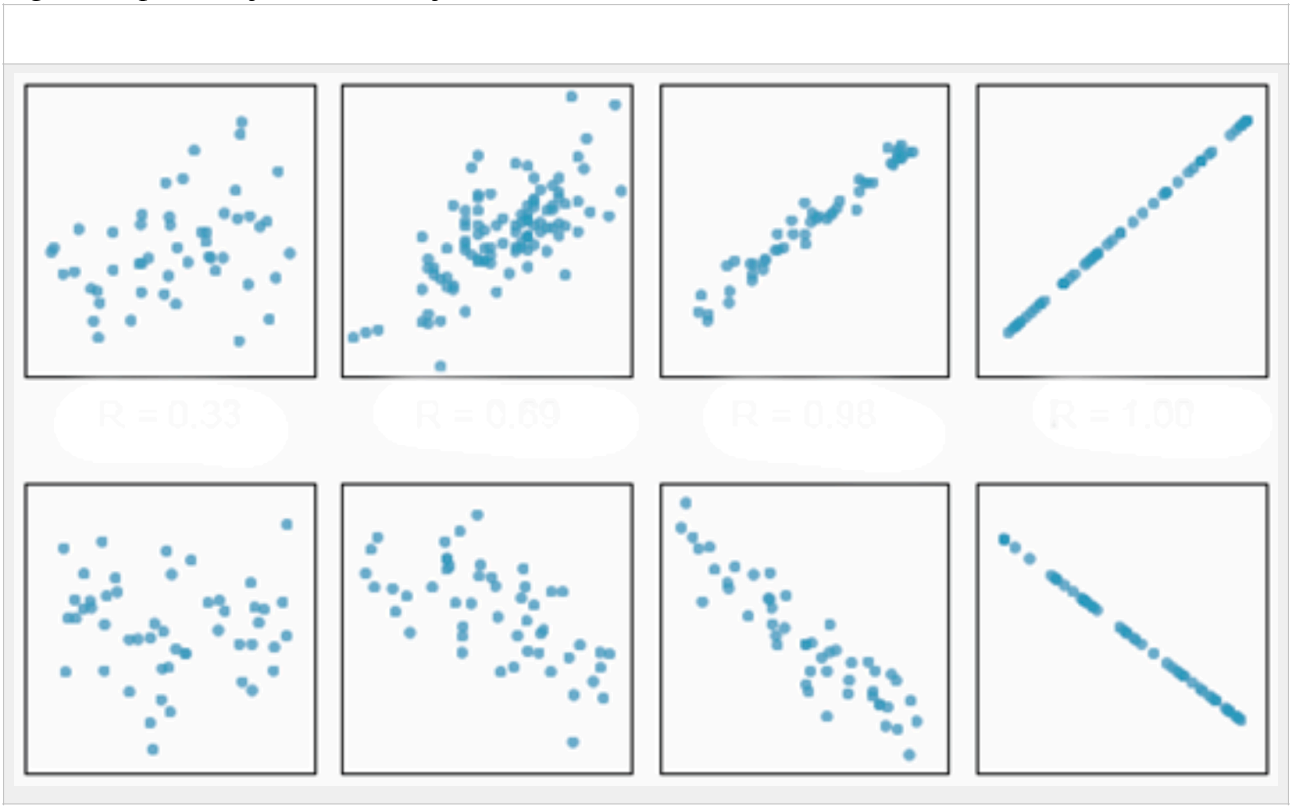
Figura: Representação de associações não lineares



Fonte: https://www.wamap.org/course/showlinkedtextpublic.php?cid=1383&id=83347

A força da associação é expressada graficamente pela proximidade dos pontos ao padrão desenhado pelos dados. Na figura seguinte, são representados gráficos de associações lineares de força crescente (positivas nos painéis superiores e negativas nos painéis inferiores).

Figura: Representação de associações lineares



Fonte: https://www.wamap.org/course/showlinkedtextpublic.php?cid=1383&id=83347

Voltando ao nosso exemplo, poderíamos dizer que a associação entre a lealdade e a duração no cargo apresenta:

- uma direção positiva;
- uma forma linear;
- de força moderada; e
- sem presença de outliers.

Correlação

Avaliar a força da relação com base no gráfico de dispersão dá-nos uma primeira aproximação que deve ser complementada com medidas numéricas, tarefa dos chamados coeficientes de correlação.1

Existem vários tipos de coeficientes de correlação, com usos diferentes. Mas todos eles (? controlar) variam entre -1 e +1, em que o valor 1 sinaliza total associação e zero a ausência de associação. Assim, um valor próximo de 1 indica uma associação forte, valores próximos de 0,50 sinalizam associações moderadas e valores abaixo de 0,30 são indicação de associações baixas ou inexistentes. O sinal do coeficiente sinaliza a direção da relação, isto é, se se trata de uma relação positiva ou negativa.

[1] É importante sublinhar que uma correlação alta entre variáveis nem sempre implica correlação. As ferramentas estatísticas que serão apresentadas aqui são apropriadas apenas para examinar relações lineares e, quando usadas em situações não lineares, podem levar a erros de raciocínio.

Coeficiente de Pearson

Existem várias medidas diferentes para o grau de correlação nos dados, dependendo do tipo de associação e o tipo de dados. O mais utilizado é o coeficiente de Pearson, adequado para avaliar a associação *linear* entre duas variáveis quantitativas. Este coeficiente é representado pela letra *r*.

A fórmula pode ser apresentada de várias formas e pretende quantificar em que medida, em média, a alteração (i.e., aumento ou diminuição) nos valores de uma variável implica alterações nos valores da outra (como as medidas de *covariância*).

Para além do valor, podemos também solicitar o teste estatístico associado às correlações - que testa se é mais ou menos provável que a correlação seja estatisticamente diferente de zero na população e quais os intervalos de confiança associados a esta estimativa (ver módulo análise inferencial).

Então, qual é a força da correlação entre lealdade e duração em funções de ministro/a?

```
cor (B$lealdade, B$duração)
```

```
[1] 0.9096987
```

A correlação positiva (r > 0) confirma que a duração no cargo de ministro geralmente aumenta com a lealdade. O valor de r indica que a relação linear é forte (r= 0.91) mas não perfeita, pelo que podemos esperar que a duração no cargo varie um pouco, mesmo entre indivíduos com o mesmo nível de lealdade.

Propriedades de r:

- A correlação não muda quando as unidades de medida de qualquer uma das variáveis mudam.
- A correlação mede apenas a força de uma relação linear entre duas variáveis.
- A correlação por si só não é suficiente para determinar se uma relação é linear ou não.
- A correlação é fortemente influenciada por outliers.

Regressão linear

Os coeficientes de -correlação fornecem uma primeira medição sobre a força e direção de uma associação. É uma medição geral e que não distingue os papéis e o impacto das variáveis (independente/dependente). Para descrever mais precisamente como uma variável afeta ou prediz o comportamento de outra, a estatística oferece as técnicas de regressão.

A mais usada é a regressão linear simples, que modela relações lineares entre duas variáveis quantitativas contínuas. A regressão linear é a técnica para encontrar a linha que melhor representa o padrão de uma relação linear.

Visualmente, tal como vimos no gráfico de dispersão, a variável independente (x) é colocada no eixo horizontal e a variável dependente (y) no eixo vertical. Desta forma, vemos como a variável y se comporta à em função dos valores x. A sua fórmula matemática é:

$$y = a + b*x$$

Em que:
- *y* é um valor da variável dependente;
- *a* é a constante de regressão;
- *b* é o declive ou coeficiente de regressão; e
- *x* é um valor para a variável independente.

O coeficiente de regressão (b) indica o valor médio de alteração da variável y associada ao aumento de uma unidade da variável x, previsto pela reta de regressão. Este coeficiente dá-nos a informação do valor que em média a variável dependente aumenta ou diminui, quando se aumenta 1 unidade na variável independente.
A constante indica o valor médio esperado para a variável dependente (y) quando a variável independente (x) é igual a zero.

PEARSON'S R VALUE	CORRELATION BETWEEN TWO THINGS IS...	EXAMPLE
r = -1	Perfectly negative	Hour of the day and number of hours left in the day
r < 0	Negative	Faster car speeds and lower travel time
r = 0	Independent or uncorrelated	Weight gain and test scores
r > 0	Positive	More food eaten and feeling more full
r = 1	Perfectly positive	Increase in my age and increase in your age

Nem sempre este valor tem interpretação fácil. Podemos centrar a variável independente (centrar uma variável é uma transformação matemática da variável em que se retira o valor da média a cada valor da variável) de forma a que o valor zero na nossa variável independente signifique o valor médio de lealdade da amostra (valor da média - valor da média = 0). Nesse caso, a constante (intercepto) indicaria o valor médio esperado para a variável dependente quando a variável independente se encontra no seu valor médio.

```
lm (B$duração~B$lealdade)
```

```
Call:
lm(formula = B$duração ~ B$lealdade)

Coefficients:
(Intercept)   B$lealdade
      77.73        129.09
```

A duração no cargo de ministro é de 77,73 dias, em média, quando o nível de lealdade ao chefe do governo é igual a zero, ou seja, quando está na média. Por cada unidade de aumento na variável independente lealdade há, em média, um aumento de 129,09 dias na variável dependente duração em funções.

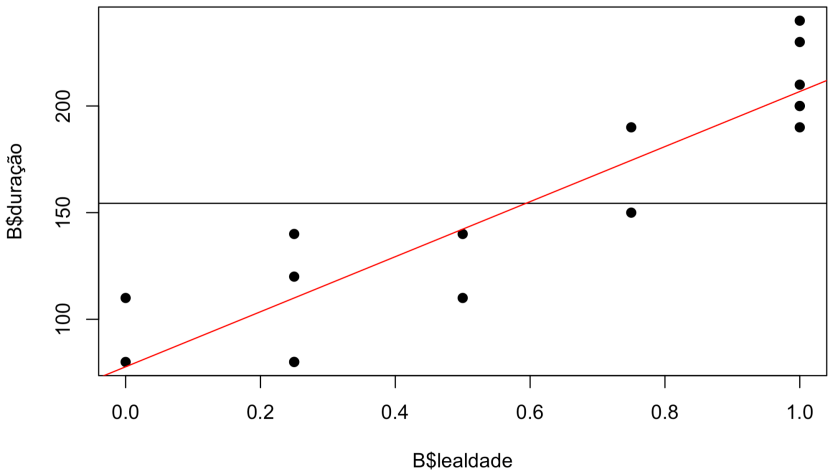
- $a = 77.73$
- $b = 129.09$

$$y = a + b*x$$
$$\text{duração} = 77.73 + 129.09 * \text{lealdade}$$

A linha de regressão, também conhecida como a linha de mínimos quadrados, é a representação gráfica do valor esperado (estimado pela fórmula) da variável y (dependente) para todos os valores da variável x (independente). É a linha de "melhor ajuste".

No gráfico de dispersão com base nas coordenadas (x, y), a linha de regressão pode ser desenhada se soubermos qual a constante (a) e o declive (b), isto é, se soubermos o ponto em que a reta intercepta o eixo vertical e a sua inclinação.

```
plot (B$lealdade, B$duração, pch=19)
abline (a=mean(B$duração), b=0)
abline(a= 77.73, b= 129.09, col="red")
```



Predição da duração em funções para um indivíduo com lealdade = 0

```
77.73 + (129.09 * 0)
```

```
[1] 77.73
```

Predição da duração para um indivíduo com lealdade = 0.25

```
77.73 + (129.09 * 0.25)
```

```
[1] 110.0025
```

Predição da duração para um indivíduo com lealdade = 0.50

```
77.73 + (129.09 * 0.50)
```

```
[1] 142.275
```

Predição da duração para um indivíduo com lealdade = 0.75

```
77.73 + (129.09 * 0.75)
```

```
[1] 174.5475
```

Predição da duração para um indivíduo com lealdade = 1

```
77.73 + (129.09 * 1)
```

```
[1] 206.82
```

Contudo, uma vez que o modelo não descreve perfeitamente os dados (a nuvem de pontos), a equação da regressão linear contém um termo de erro:

$$y = a + b*x + erro$$

O erro padrão da estimativa é o desvio padrão dos pontos de dados conforme eles são distribuídos em torno da linha de regressão. O erro padrão pode ser usado para desenvolver intervalos de confiança em torno de uma previsão.

4. Teste Estatístico

```
summary(lm (B$duração~B$lealdade))
```

```
Call:
lm(formula = B$duração ~ B$lealdade)

Residuals:
   Min     1Q Median     3Q    Max
-32.27 -18.75   0.00  17.39  33.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    77.73      11.05   7.035 5.91e-06 ***
B$lealdade    129.09      15.75   8.197 1.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.54 on 14 degrees of freedom
Multiple R-squared:  0.8276,    Adjusted R-squared:  0.8152
F-statistic: 67.18 on 1 and 14 DF,  p-value: 1.031e-06
```

Em primeiro lugar, vemos novamente que se um ministro tiver um nível de lealdade para com o chefe do governo igual a 0, esse ministro durará 77,73 dias, em média, em funções. Além disso, caso um ministro tenha um nível de lealdade igual a 1, esse ministro aumenta a sua duração no cargo em 129,09 dias.
O erro padrão é uma estimativa do desvio padrão do coeficiente que nos diz quanta incerteza está associada ao nosso coeficiente. Podemos, por exemplo, construir um intervalo de confiança à volta do nosso slope (lealdade):
129,09 ± 1,96 (15,75) = [98,22; 159,96].
Podemos assim dizer, com 95% de confiança, que o declive da reta está entre 98,22 e 159,96 dias.
O t-value e o p-value ajudam-nos a perceber quão significativo é o nosso coeficiente para o modelo. Neste caso, vemos que o valor-p para o intercepto e para o declive (i.e., lealdade) são extremamente pequenos, o que significa que estes dois coeficientes ***não são*** *zero, ou seja, estes coeficientes acrescentam valor ao modelo e ajudam a explicar (parte da) variância da nossa variável dependente. Por fim, o Multiple R-squared diz-nos a percentagem de variância da variável dependente que é explicada pela variável independente. Neste exemplo, a lealdade explica cerca de 82,76% da variação no número de dias em funções.*

5. EXERCÍCIOS

Secção PANORAMA

1) Identificar e escolher 5 artigos vinculados a um tópico do seu interesse, que utilizem análise estatística.
2) Para cada um dos 5 artigos escolhidos, identificar que tipo de análise aplicam (descritivo, explicativo) e que técnicas e ferramentas (coeficientes, regressões, caixa de bigodes, etc.).
3) Para 2 dos 5 artigos escolhidos,
   a) identificar e descrever a base de dados utilizada (unidade de observação, número de observações, variáveis) e
   b) reconstruir o codebook incluindo pelo menos 5 variáveis.

Secção UMA QUANTITATIVA

4) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável numérica. Transcrever essa descrição.

Secção UMA QUALITATIVA

5) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável qualitativa. Transcrever essa descrição.

Secção DUAS QUANTITATIVAS

1) Para 1 dos 5 artigos escolhidos, identifique a variável independente e a variável dependente. Classifique-as (e.g., categórica, numérica).
2) Agora escolha um artigo com uma variável independente numérica e uma variável dependente numérica. Calcule o coeficiente de correlação entre as duas variáveis.
3) Por fim, calcule um modelo de regressão linear e interprete os resultados.

6. Referências

Diez et al (2021). OpenIntro Statistics. https://stats.libretexts.org/@go/page/270

Bryman et al. (2021). Bryman's Social Research Methods, Oxford University Press.

APÊNDICE (R)

Gráficos com Pacote Base

Plot com linha de regressão

Plot com linha de regressão e 'loess fit'

Gráficos com pacote ggplot2

Plot

Plot com linha de regressão

Plot com linha de regressão e intervalo de confiança

cores

Com "loess smoothed fit curve"

Modificando forma pontos

Points shapes available in R

Multi-plots

Cores e formas por categoria

Multi-paneis

Texto nos pontos

'Bubble chart'

Gráficos com pacote ggpubr