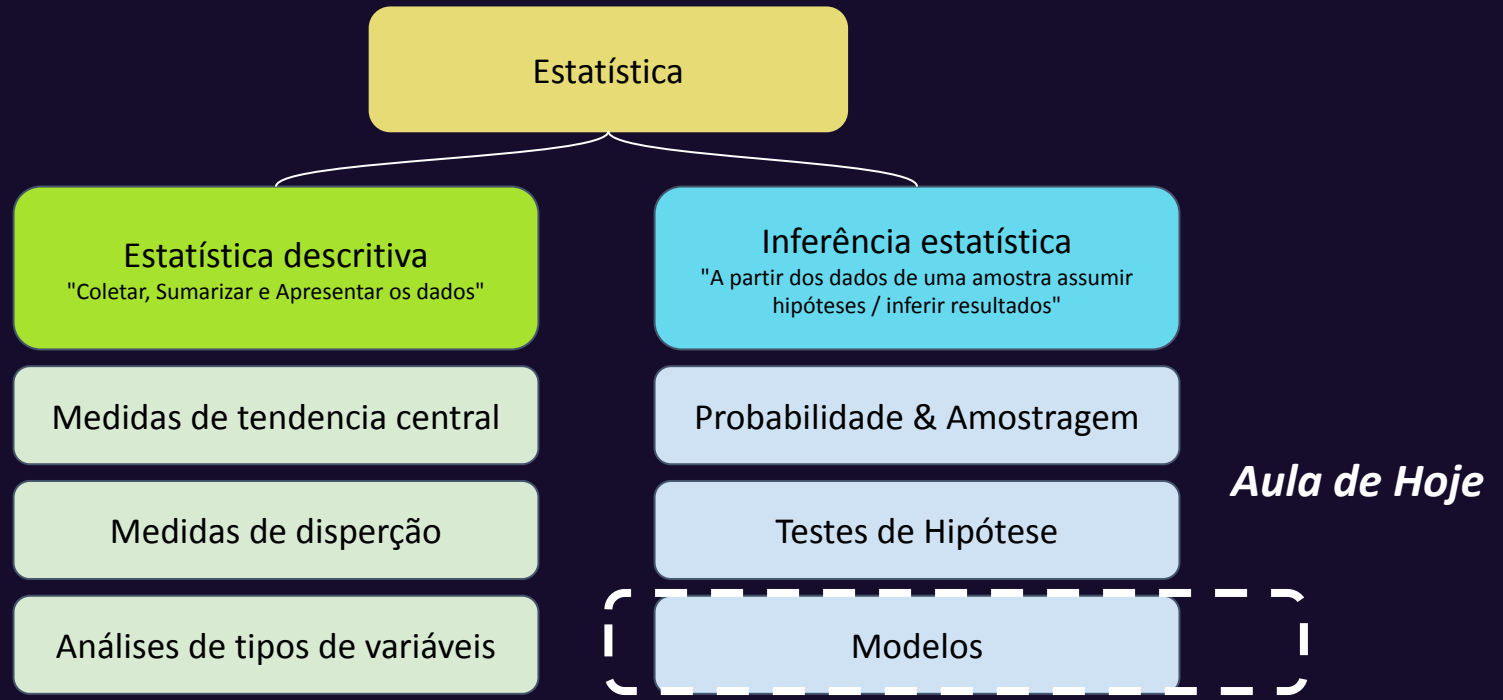
Bootcamp Data Analytics

Correlação e Regressão

WOMAKERS CODE

Na Aula de hoje vamos estudar ferramentas que potencializam a analise de dados: Correlação & Regressão



Estatística : Correlação e Regressão

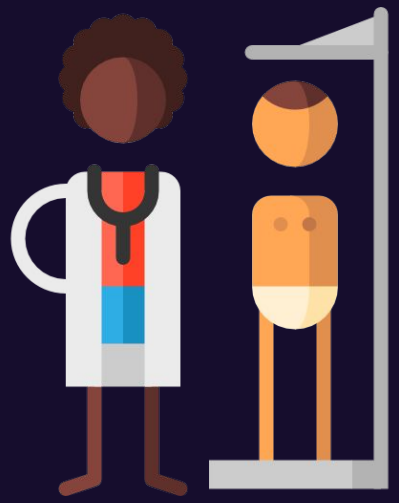
Conceitos

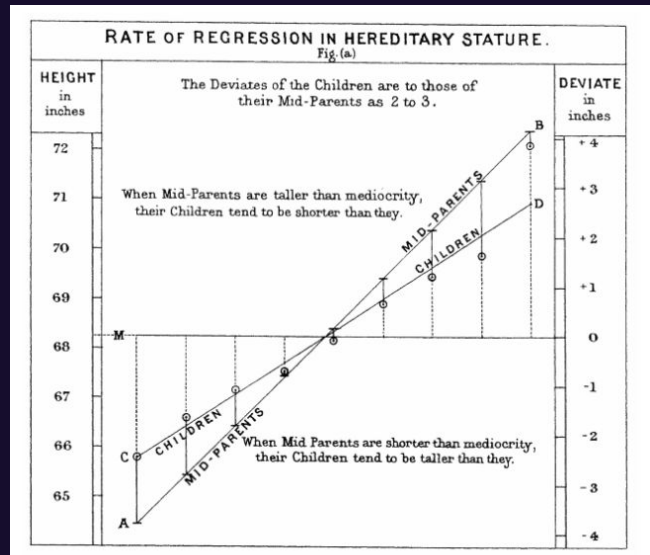
WOMAKERS CODE

Correlação: Origem

- Em 1888 um cientista chamado Francis Galton estava analisando dados da altura de crianças e seus pais em um hospital de londres.

- Galton percebeu que a altura da criança seguia uma tendência quando analisada conjuntamente com os dados dos pais. Logo a altura do pai e a altura da criança pareciam seguir uma tendencia.





from *Regression towards mediocrity in hereditary stature* (1886)

O conceito de correlação

Mas o que é correlação ? É uma medida estatística que mede o grau da força da relação entre o movimento de duas variáveis.

Ela pode assumir valores de -1 a +1 de modo que:

1. Quanto mais próximo de 1 as variáveis são mais relacionadas positivamente. Se X aumenta Y aumenta.
2. Quanto mais próximo de -1 as variáveis são mais relacionadas negativamente. Se X aumenta Y diminui
3. Quanto mais próxima a zero. Mais as variáveis não caminham juntas. Não existe uma tendência na relação entre X e Y.

Para analisar a relação das variáveis podemos utilizar o gráfico de dispersão (scatterplot em ingles)

Correlação e a dispersão das variáveis

Podemos visualizar a correlação pelo gráfico de dispersão de modo que, variáveis positivamente correlacionadas seguem uma linha de tendência para cima (Gráfico 1.) Variáveis negativamente correlacionadas seguem uma linha de tendência para baixo (Gráfico 2) e Variáveis não correlacionadas não seguem uma tendência definida.

Gráfico 1: Correlação Positiva

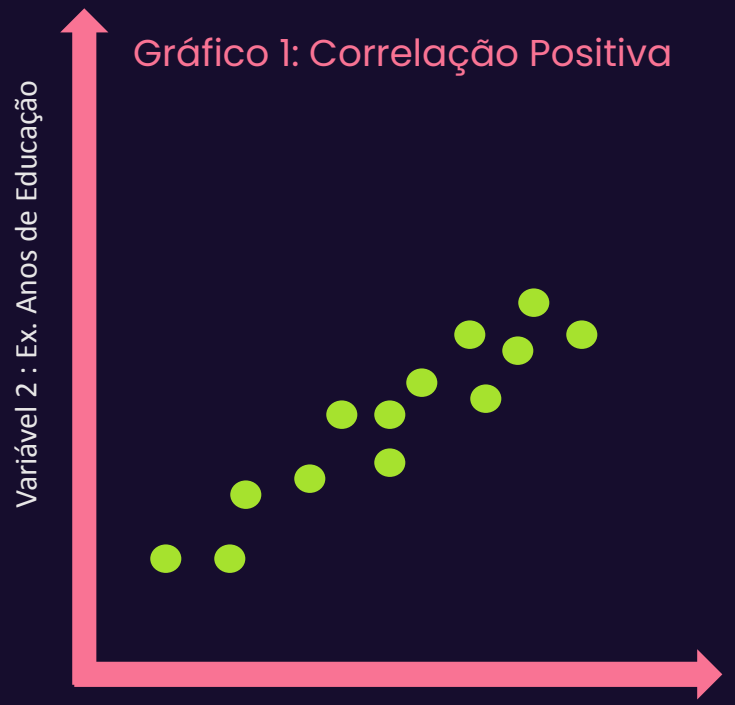


Gráfico 2: Correlação Negativa

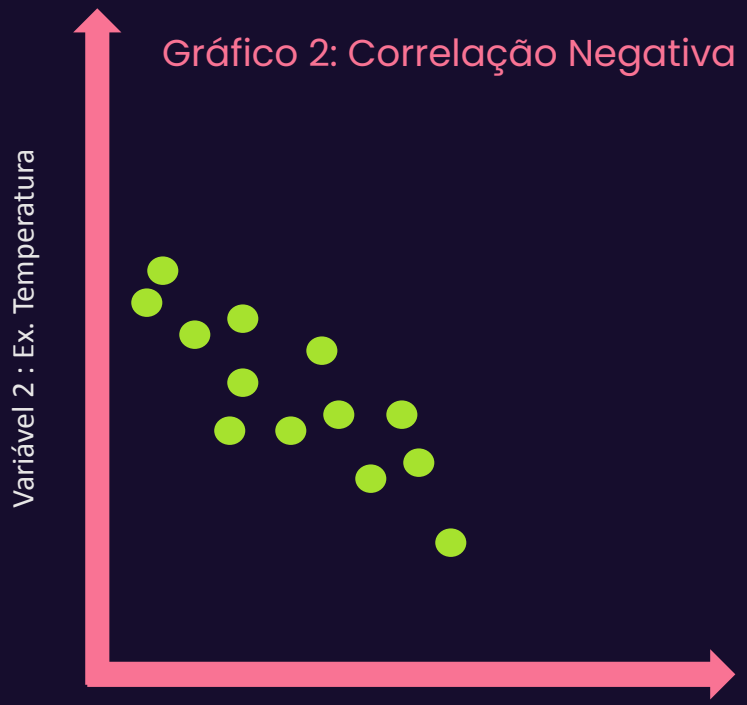
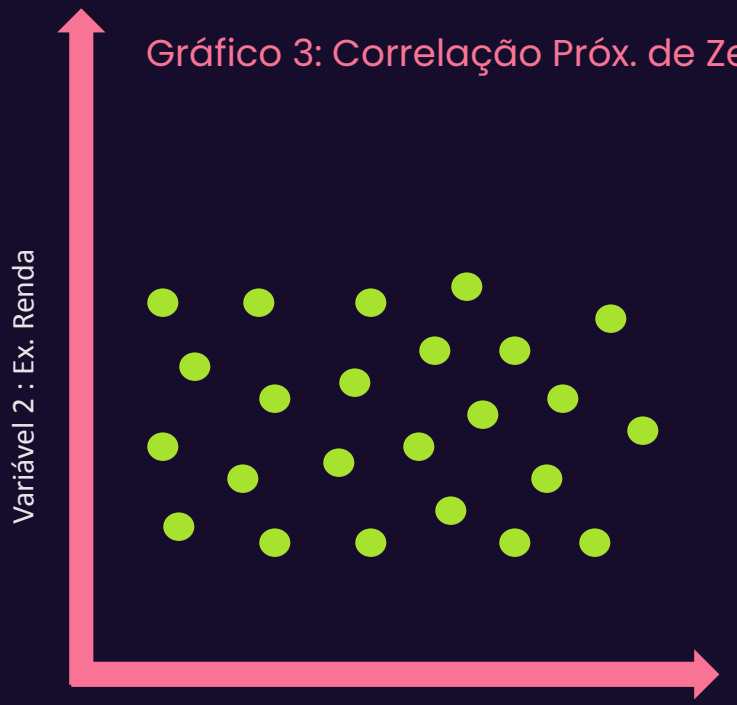


Gráfico 3: Correlação Próx. de Zero



Fórmula da Correlação

A correlação será a covariância das variáveis X e Y dividida pelo desvio padrão de x que multiplica o desvio padrão de y.

Mas o que é covariância? Ex: Queremos medir a covariancia de renda e anos de educacao.

1. Fazemos uma pesquisa como n = 1,000 pessoas e anotamos a renda, representada por X e educacao representada por Y.
2. Calculamos a média da renda, representada por $\bar{x}$ e a média da educacao representada por $\bar{y}$
3. Para cada pessoa i, vamos calcular (renda - média da renda)*(educação - média da educacao) e vamos somando até a pessoa de número n = 1,000.

$$\text{Correlacao}(X,Y) = \frac{\text{Covariância}(X,Y)}{\sigma_x \sigma_y} = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)}{\sigma_x \sigma_y}$$

Calculando a covariância para n = 5

Exemplo : Vamos calcular a correlação entre renda e educação para 5 pessoas dados os dados coletados abaixo:

Renda (X)	Educação em anos (Y)
3,000	12
6,000	17
7,000	19
4,000	12
15,000	22

1.Covariancia das variáveis:
Média de x : 7,000 ; Média de Y: 16.4

indivíduo 1: (3,000 - 7,000)*(12-16.4) = 17,600
indivíduo 2: (6,000 - 7,000)*(17-16.4) = -600
indivíduo 3: (7,000 - 7,000)*(19-16.4) = 0
indivíduo 4: (4,000 - 7,000)*(12-16.4) = 13,200
indivíduo 5: (15,000 - 7,000)*(19 - 16.4) = 44,800

soma: 75,000

2. Desvio padrão das variáveis:

Renda
indivíduo 1: (3,000 - 7,000)^2 = 16,000,000
indivíduo 2: (6,000 - 7,000)^2 = 1,000,000
indivíduo 3: (7,000 - 7,000)^2 = 0
indivíduo 4: (4,000 - 7,000)^2 = 9,000,000
indivíduo 5: (15,000 - 7,000)^2 = 64,000,000
soma = 90,000,000
desv.p = sqrt(soma) = 9487

idade:
indivíduo 1: 19.36
indivíduo 2: 0.36
indivíduo 3: 6.76
indivíduo 4: 19.36
indivíduo 5: 31.36
soma = 77,20
desv.p = sqrt(77.20) = 8.78

Calculando a covariância para n = 5

Exemplo : Vamos calcular a correlação entre renda e educação para 5 pessoas dados os dados coletados abaixo:

Renda (X)	Educação em anos (Y)
3,000	12
6,000	17
7,000	19
4,000	12
15,000	22

1.Covariancia das variáveis:
Média de x : 7,000 ; Média de Y: 16.4

indivíduo 1: (3,000 - 7,000)*(12-16.4) = 17,600
indivíduo 2: (6,000 - 7,000)*(17-16.4) = -600
indivíduo 3: (7,000 - 7,000)*(19-16.4) = 0
indivíduo 4: (4,000 - 7,000)*(12-16.4) = 13,200
indivíduo 5: (15,000 - 7,000)*(19 - 16.4) = 44,800

cov = soma: 75,000

$$Corr(X,Y) = \frac{Covariância = 75,000}{\sigma x\ \sigma y = 9487*8.78} = 0.8997$$

2. Desvio padrão das variáveis:

Renda
indivíduo 1: (3,000 - 7,000)^2 = 16,000,000
indivíduo 2: (6,000 - 7,000)^2 = 1,000,000
indivíduo 3: (7,000 - 7,000)^2 = 0
indivíduo 4: (4,000 - 7,000)^2 = 9,000,000
indivíduo 5: (15,000 - 7,000)^2 = 64,000,000
soma = 90,000,000
desv.p x = sqrt(soma) = 9487

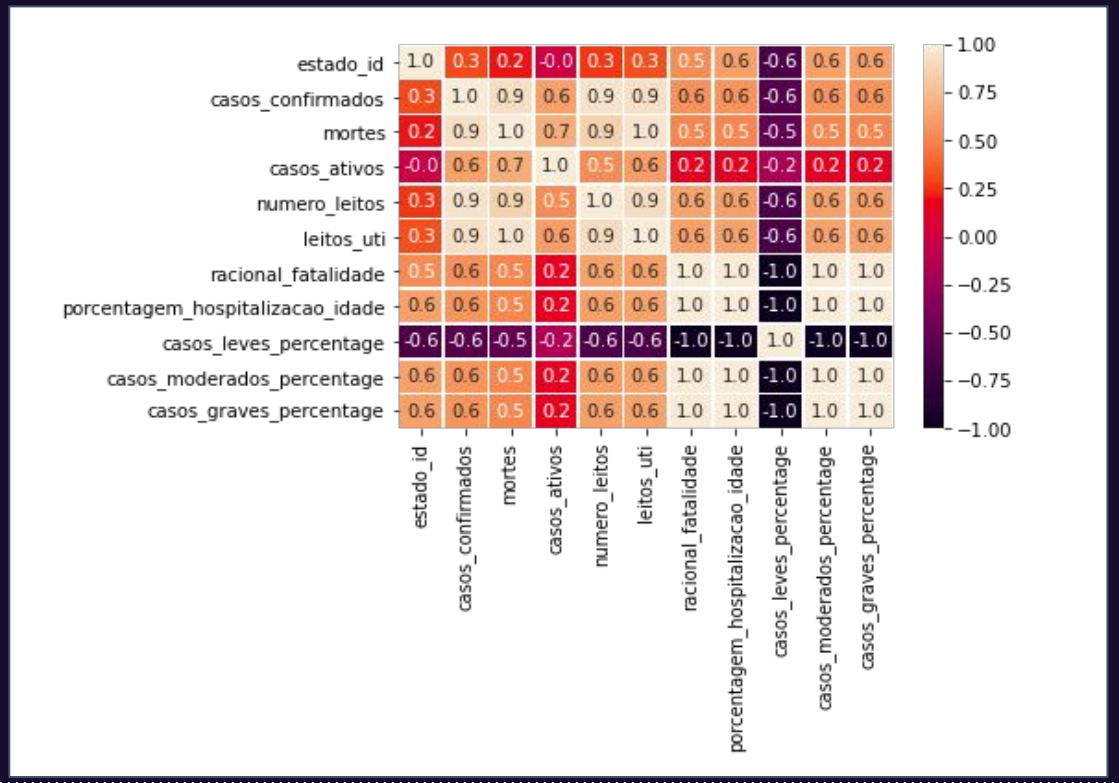
idade:
indivíduo 1: 19.36
indivíduo 2: 0.36
indivíduo 3: 6.76
indivíduo 4: 19.36
indivíduo 5: 31.36
soma = 77,20
desv.p y = sqrt(77.20) = 8.78

Exemplos de uso

- Na prática usamos muito a correlação quando queremos ver a relação de duas variáveis numéricas.

- Mas como analisar quando temos muitas variáveis numéricas? Podemos fazer uma matriz de correlação

- Ex: Matriz de correlação:
  COVID 19



O que fazer quando temos variáveis categóricas?

Podemos medir a correlação somente de variáveis numéricas, mas suponha que você queira analisar a correlação de: região e incidência de covid, como podemos proceder?

Muito de fala em dumização, que seria o processo de atribuir um numero a uma variável categórica e mensurar a correlacao a partir deste numero. Ex: sul = 1, sudeste =2,norte = 3 ...

Porém não podemos analisar a correlação nesse caso. Isso porque estaríamos atribuindo um peso a uma determinada regiao, assumindo que a região 3 > região 2, e assim por diante.

O que fazer quando temos variáveis categóricas?

Sendo assim, nesse caso não podemos analisar a correlação. alternativas:

"Correlação" entre uma numérica e uma categórica:

1. Análise visual do boxplot da variável numérica, separado por cada uma das categorias.
2. Teste anova: podermos fazer um teste anova para ver se a média de da variável numérica e estatisticamente diferente para as diferentes categorias.

O que fazer quando temos variáveis categóricas?

Sendo assim, nesse caso não podemos analisar a correlação. alternativas:

"Correlação" duas variáveis categóricas:

1. Podemos criar a tabela de contingências em que vemos as frequências de ocorrências nas combinações das diferentes categorias.

2. Podemos fazer a análise de correspondência utilizando o teste qui-quadrado. Ele compara as contagens de frequência observadas na tabela de contingência com as contagens de frequência esperadas.

Estatística : Correlacao & Regressao

Correlacao vs Causalidade

WOMAKERS CODE

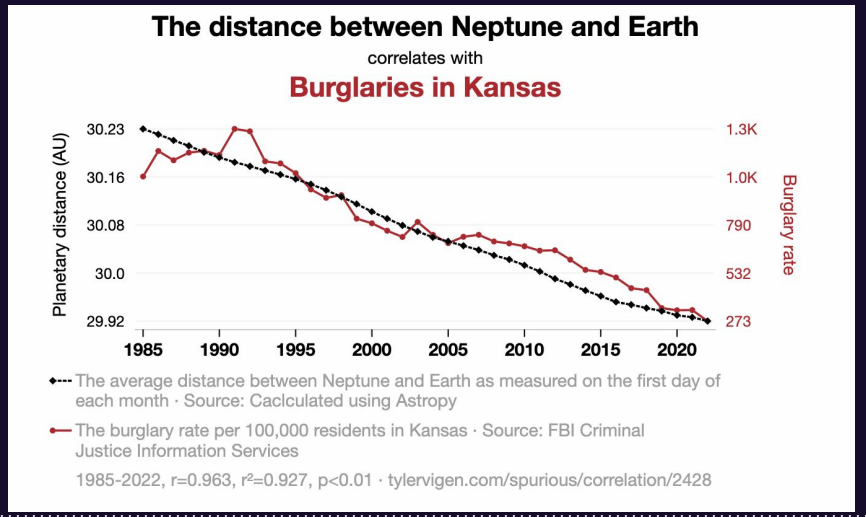
Correlação forte não necessariamente implica em causalidade.

Duas variáveis podem ser altamente correlacionadas, porém uma não causar a outra. Isso pode ocorrer por alguns motivos, dentre eles:

1. A existência do que chamamos de correlação espúria.

   Uma correlação espúria é uma correlação que existe simplesmente ao acaso, sem sentido algum.

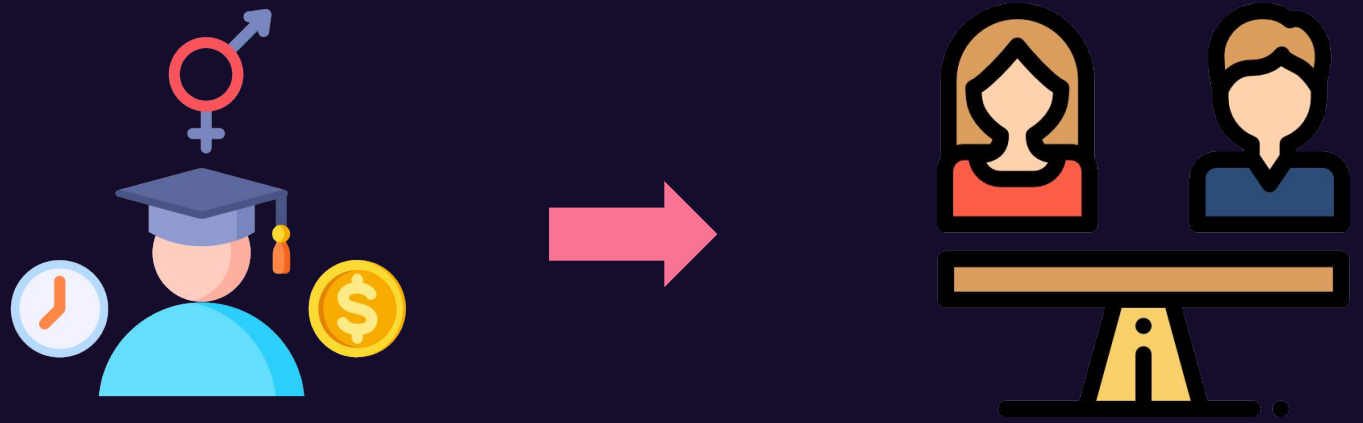
Exemplo: Assaltos no Kansas
e a distância
entre Netuno e a Terra



Correlação forte não necessariamente implica em causalidade.

2. Viés de variável omitida: Pode acontecer de variáveis serem altamente correlacionadas, por exemplo genero e renda. Mas será mesmo que o genero causa menor renda? ...

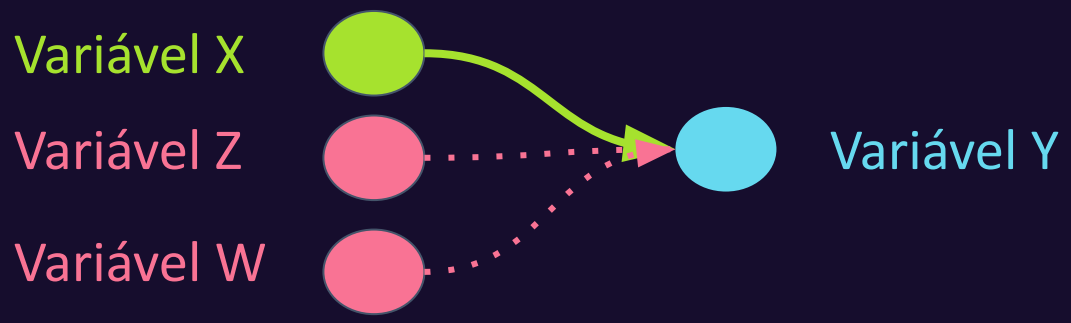
Na verdade a razão muitas das vezes por mulheres obterem menor renda vem do baixo incentivo a educação, que no caso se correlaciona diretamente com renda, sendo a educacao a "variável omitida".



E como proceder para exprimir causalidade?

Modelos regressivos podem nos ajudar a exprimir a causalidade que não conseguimos enxergar na correlação.

Isso porque atraves desses modelos que iremos ver, poderemos incluir diversas variáveis que estao omitidas na correlacao. limpando o viés de variável omitida na relação entre X e Y.



Atraves dos modelos de regressão podemos ver o efeito de X em Y limpando o viés das variáveis omitidas Z e W

Estatística : Correlacao & Regressao

Modelos de Regressão

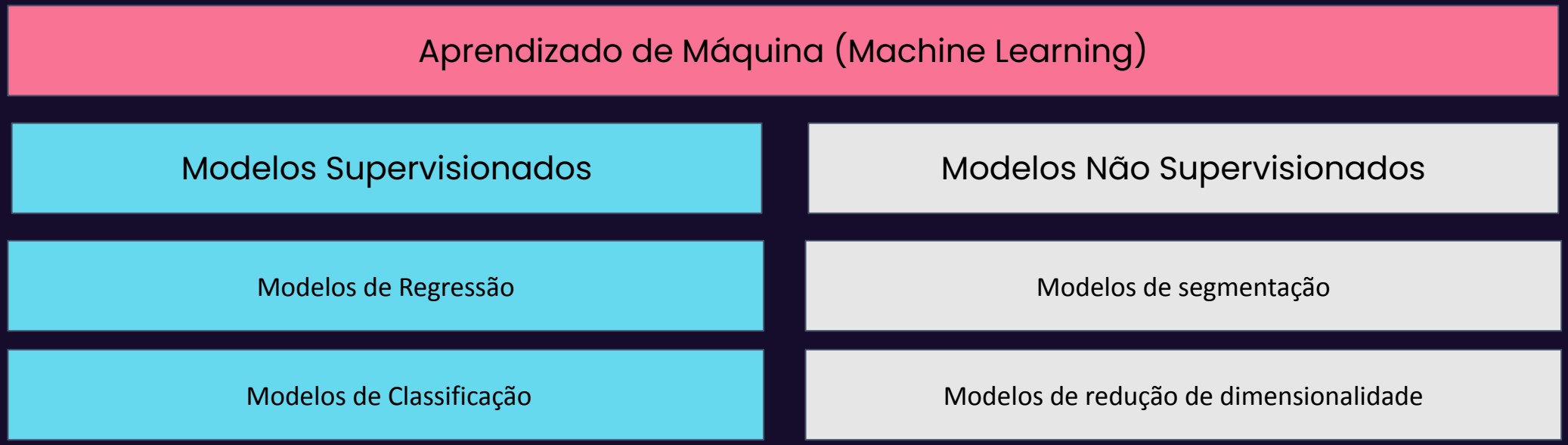
WOMAKERS CODE

Introdução ao Machine Learning

Aprendizado de Máquina (Machine Learning)

Modelos Supervisionados	Modelos Não Supervisionados
São modelos em que temos uma variável de interesse que desejamos **prever seu valor**. Essa variável de interesse, pode ser chamada de variável target ou variável dependente. Ex: preço, renda...	São modelos utilizados na detecção de padrões. Nesse caso o próprio algoritmo encontra os padrões e não temos uma variável target.

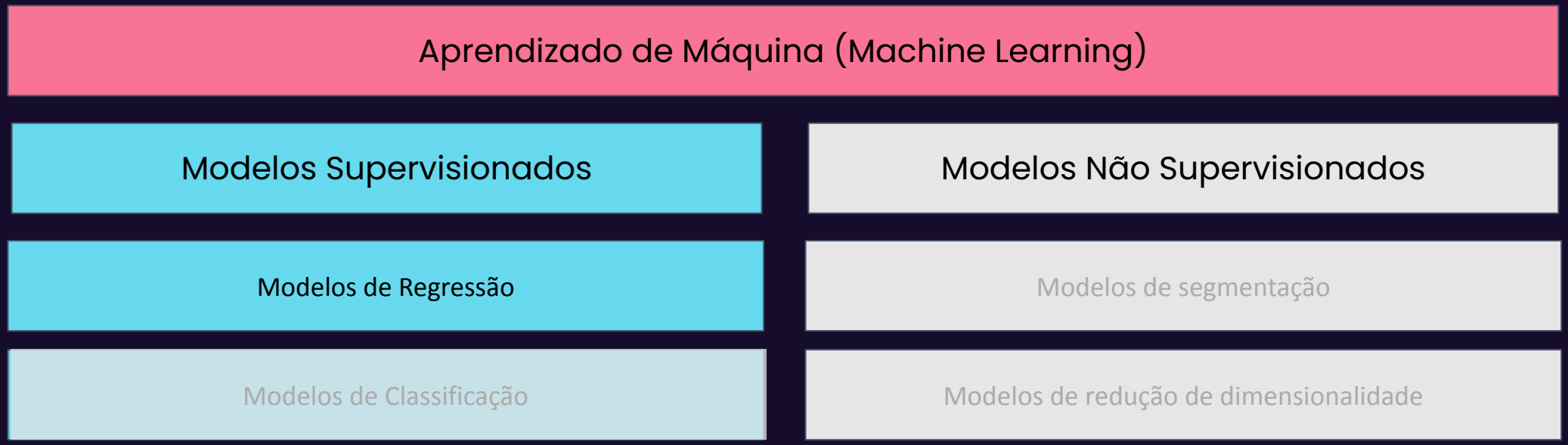
Supervisionados vs Não Supervisionados



Os modelos supervisionados podem ser divididos em:

1. modelos de regressão nos quais a variável target, que queremos prever sao contínuas. Exemplo: renda, preço de um imóvel.
2. modelos de classificação nos quais a variável target, que queremos prever são binárias (podem ser chamadas também de variáveis dummy), assumindo valores de 0 ou 1. Exemplo: fraude (1) ou não fraude (0), inadimplência (1) ou adimplência

No universo dos modelos supervisionados



Na aula de hoje iremos introduzir os modelos de regressão. focando em um tipo de modelo específico denominado regressão linear.

Regressão Linear

A regressão linear simples advém da estatística, muito antes da "descoberta do machine learning".

Ela também foi descoberta por Francis Galton, assim como a correlação nos seus estudos da associação da altura de crianças e seus pais.

Chamamos de regressão linear simples um modelo em que temos:

1. A variável target: Y, na qual queremos prever
2. Uma variável explicativa X, que explica o comportamento da variável Y.
3. A relação entre X e Y é linear.
4. Um nível de erro e, que é o nível de erro do modelo.

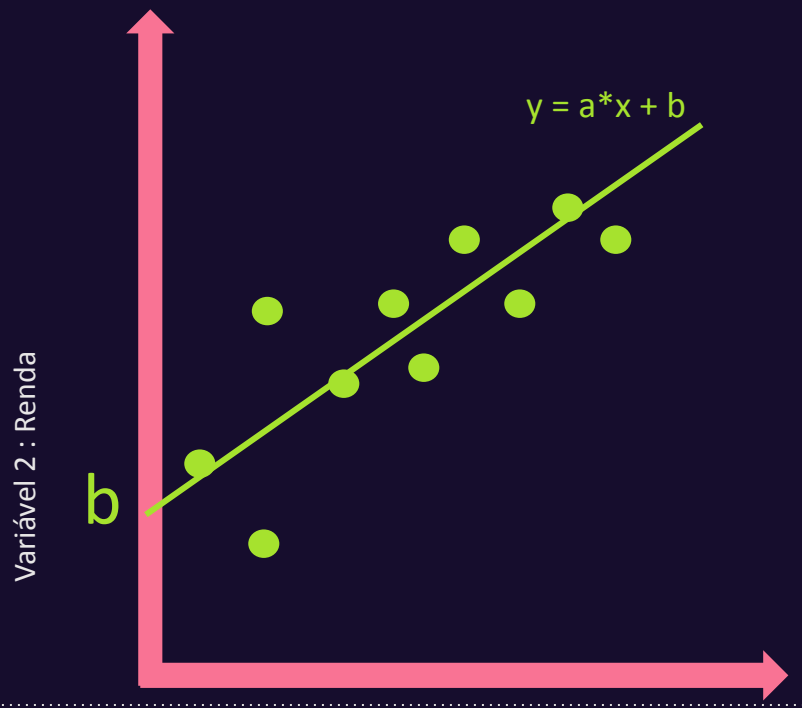
Sendo assim, podemos utilizar a equação de uma reta, para descrever a relação de X e Y:

$Y_i = aX_i + b + e$

Regressão Linear

Relembrando os conceitos de equações lineares a regressão simples nada mais é do que a equação que explica a relação de X e Y.

De modo que de posse do coeficientes angular a e do coeficiente linear b, podemos prever valores de Y!!!

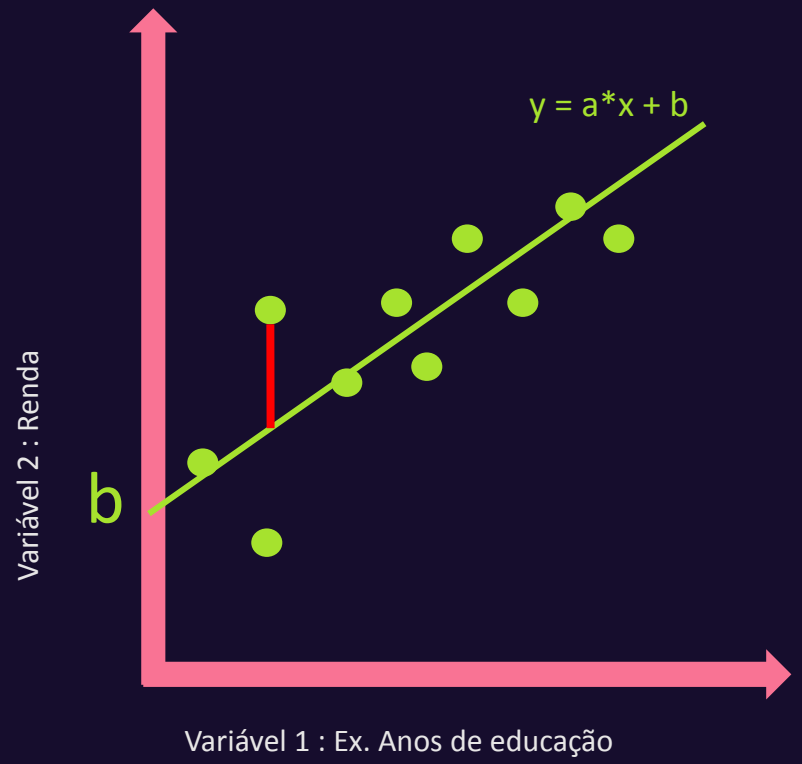


Sendo assim, o grande objetivo desse modelo é estimar os coeficientes a e b, de modo a obtermos o menor erro possível.

Exemplo: Queremos estimar a renda de uma pessoa a partir dos anos de educação dela.

Como estimar os coeficientes?

Existem diversas formas de estimar o valor de a e b na equação da reta. Mas um dos métodos mais comuns é atraves do método MQO "minimização dos quadrados ordinários"

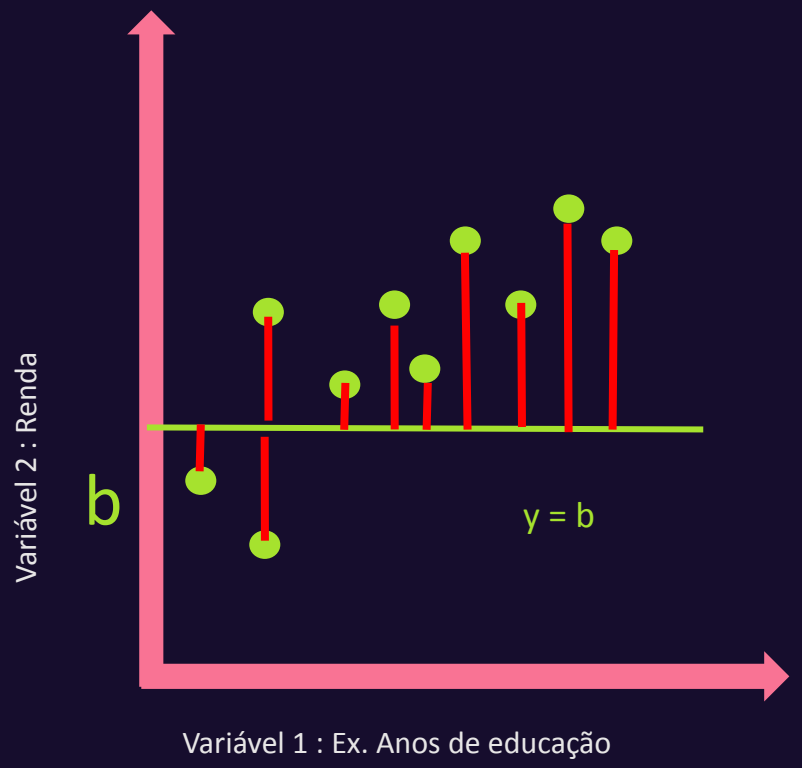


Para cada observação da base de dados temos um pontinho no gráfico de dispersão e a distância de cada pontinho a uma determinada reta (em vermelho) é chamada de resíduo.

No método MQO, vamos encontrar valores de a e b que minimizem a soma dos quadrados dos resíduos.

Na prática como esse método funciona?

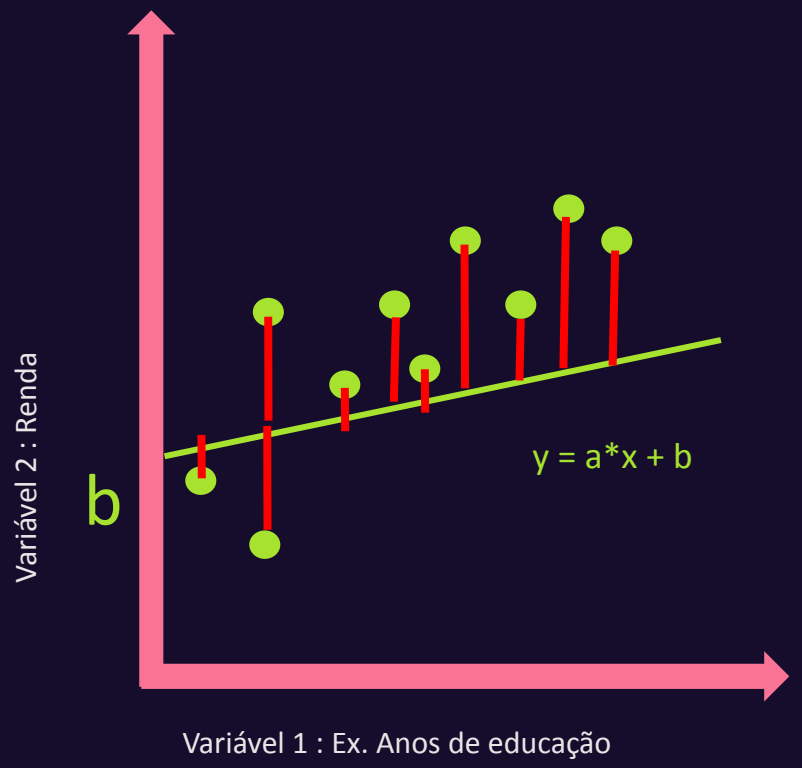
Visualizando a regressão linear



Construímos uma reta aleatória aqui e vimos a distância de cada pontinho até essa reta (o resíduo). elevamos esses valores ao quadrado e somamos esses valores, obtendo a soma dos quadrados os resíduos.

... vamos tentar rotacionar essa reta agora

Visualizando a regressão linear

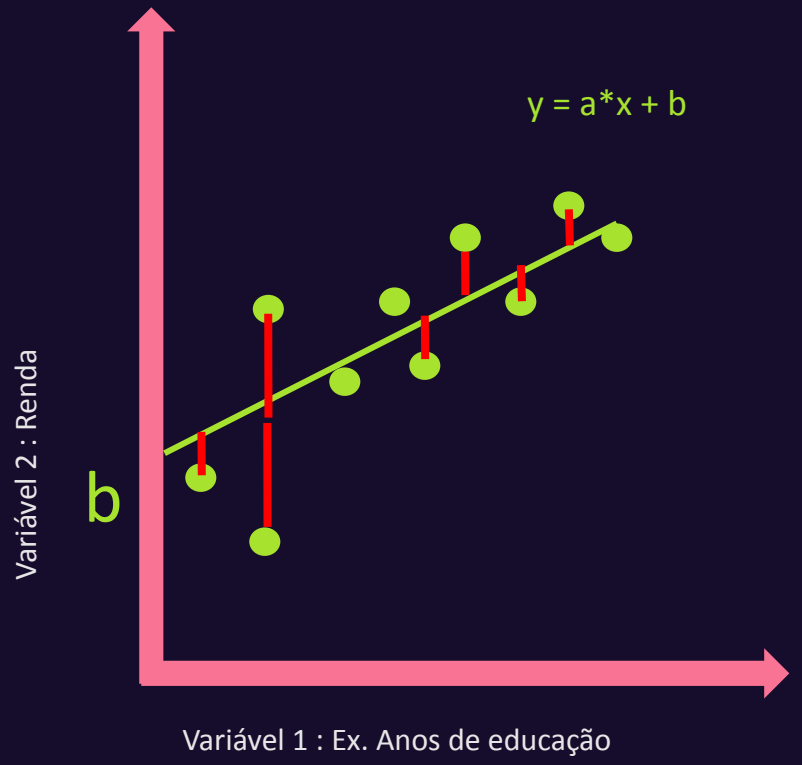


Rotancionando a reta e recalculando a etapa anterior vemos que obtemos valores menores de SQR (retas menores)

e o ajuste melhora!

vamos ajustar um pouco mais! até que esse valor seja o menor possível

Visualizando a regressão linear



Pronto agora obtivemos os menores valores possíveis da soma dos quadrados dos resíduos.

Os valores de a e b, que minimizam essa soma, serão os parâmetros estimados!!

Na prática, podemos encontrar esses valores resolvendo sistemas de equações matriciais.

Em python iremos utilizar o pacote statsmodels.

Nomenclaturas: Erro ou Resíduo

- Erro: É a diferença de um dado observado real de um valor de interesse.
- Resíduo: É a diferença de um dado observado e valor estimado.
- Vamos supor que queremos prever o valor da renda atraves da variável anos de estudo.

$$Renda = c + a*Anos\ de\ estudo + e$$

O e na equação acima será o erro do modelo. Um desvio que assumimos que existe.

Suponha que chegamos em $c = 1000$ e $a = 250$ através do método MQO.

Nomenclaturas: Erro ou Resíduo

Assim:

$$\hat{R} = 1000 + 250*X1 + e$$

Onde chamamos de $\hat{R}$ a renda estimada ; e X1 os anos de educação.

Suponha agora que coletamos uma observação que coletamos de: Joana (3900 de renda , 14 anos de estudo)

Assim a renda estimada de joana aplicando no modelo será:

$\hat{R} = 1000 + 250*14 = 4500$

Assim o resíduo será: Observado- $\hat{R}$ - = 3900- 4500
E o salário de Joana está abaixo do estimado.

Output de um modelo regressivo

Para resolver um problema de regressão linear precisamos como input colocar as variáveis X e Y, sendo elas variáveis numéricas.

E os pacotes irão nos retornar:

1. Os coeficientes a e b

2. Uma tabela com o Teste t realizado na variável x: esse teste tem como hipótese nula que o coeficiente angular da regressão é igual a zero. de modo que se o p-valor < 0.05 significa que existe uma relação linear entre X e Y!!!

3. O coeficiente de "fit" do modelo, conhecido por R- quadrado. O R-quadrado será uma medida de 0 a 1, de modo que quanto mais próximo de 1, significa que mais os dados seguem uma relação linear perfeita.

A tabela de regressão

A tabela ao lado sumariza os outputs mencionados.

O Campo coef indica o valor do coeficiente. nesse caso, a equação da reta é:

y = -3.2002 + 0.7529*x1 + e

"A cada 1 unidade de x1 y aumenta em 0.7529 unidades - 3.20"

O p-valor é menor que 0.05 o que indica que a regressão é válida!!

E o r-quadrado (R-squared) é 0.669 um valor satisfatório.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.669
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	299.2
Date:	Mon, 01 Mar 2021	Prob (F-statistic):	2.33e-37
Time:	16:19:34	Log-Likelihood:	-88.686
No. Observations:	150	AIC:	181.4
Df Residuals:	148	BIC:	187.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.2002	0.257	-12.458	0.000	-3.708	-2.693
x1	0.7529	0.044	17.296	0.000	0.667	0.839

Omnibus:	3.538	Durbin-Watson:	1.279
Prob(Omnibus):	0.171	Jarque-Bera (JB):	3.589
Skew:	0.357	Prob(JB):	0.166
Kurtosis:	2.744	Cond. No.	43.4

R-Quadrado

O R-quadrado é um dos principais outputs do modelo de regressão e ele pode ser calculado da seguinte maneira:

Sua interpretação será: qual o % da variabilidade dos dados explicado pelo modelo

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Sendo Ŷ o valor de y estimado utilizando a equação da reta encontrada!

A interpretação do seu valor se dá da seguinte forma:

1. R^2 > 0.6 -> relação forte
2. R^2 < 0.6 e R^2 > 0.3 -> relação moderada
3. R^2 < 0.3 -> relação fraca

Overfit

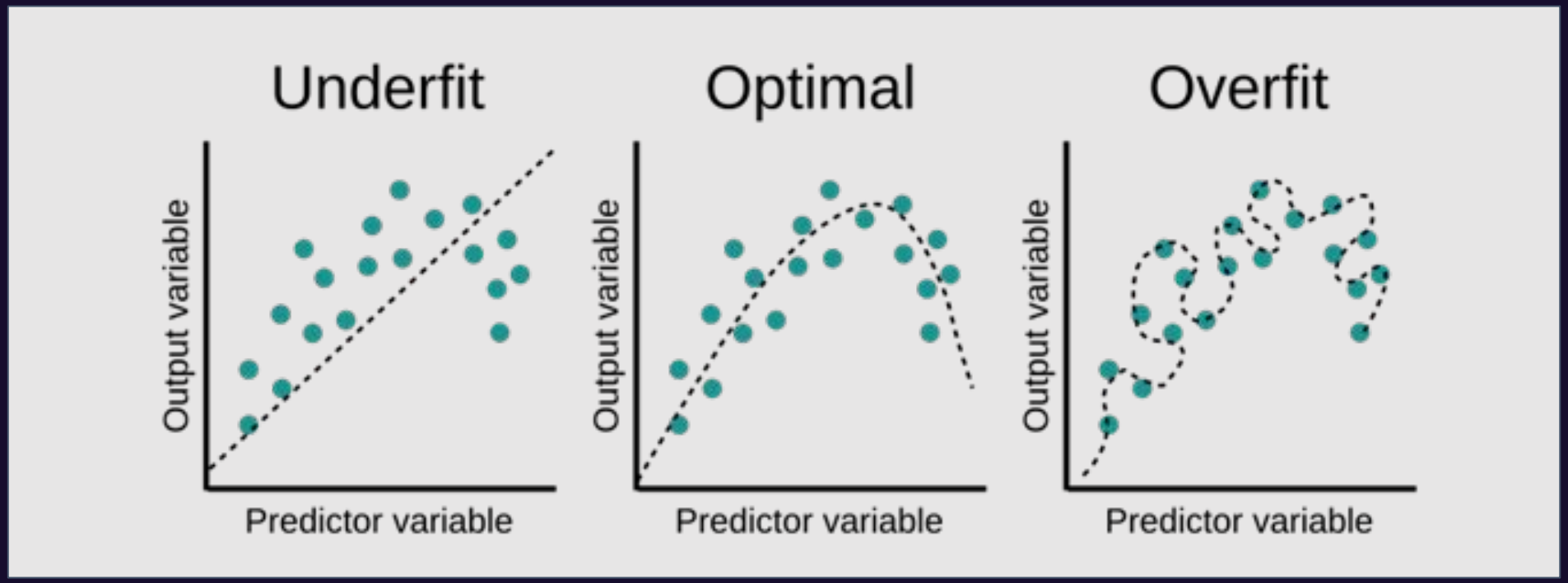
Quando o R-quadrado possui valor muito elevados, próximo a 1, pode não ser um bom sinal. :(

Pois podemos ter um indício de um dos problemas frequentes na análise de dados, o chamado overfit.

O overfit acontece quando a relação descrita entre X e Y é tão tão perfeita que o modelo perde generalização

Overfit

Visualmente, é como se um modelo fosse perfeitamente ajustado aos dados do input de modo que o erro dele quando surgir uma nova informação será elevado.



Vies de variável omitida

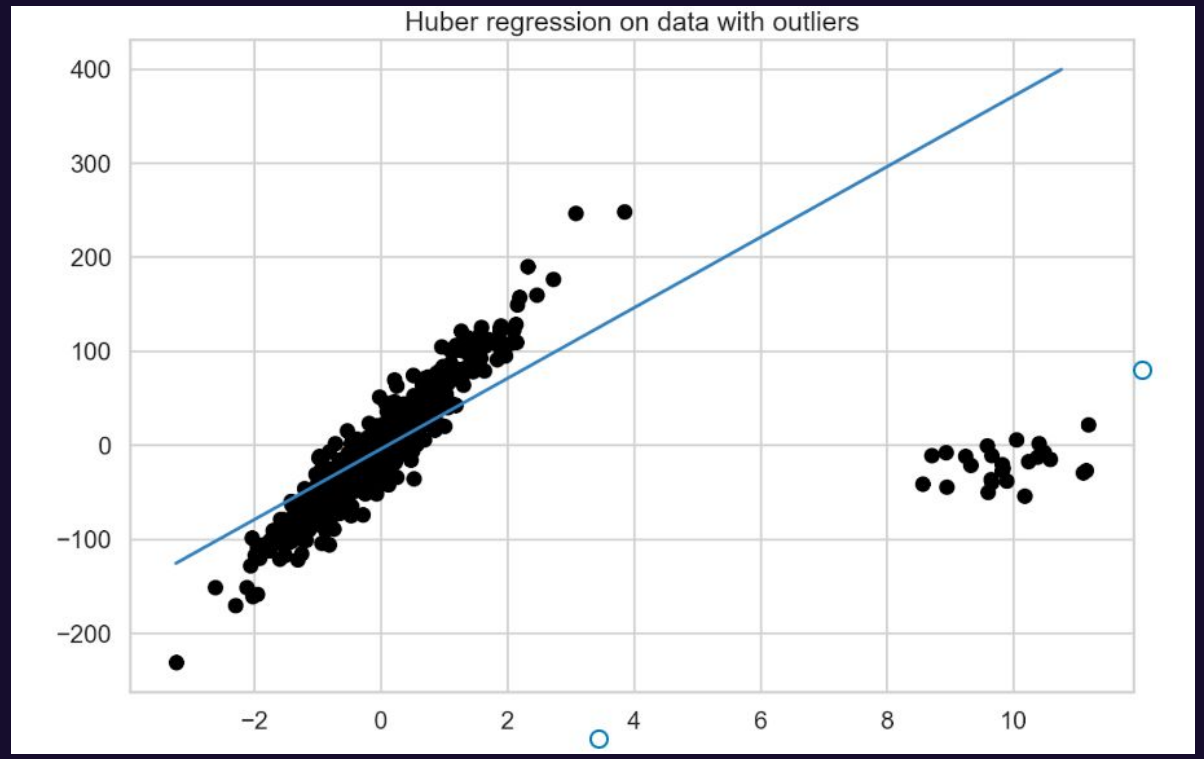
No caso do R quadrado elevado na regressão simples, ele pode refletir não somente o overfit mas também uma correlação espúria, ou um viés de variável omitida, como vimos na correlação.

Para corrigir esse problema, podemos adicionar mais variáveis explicativas no modelo, deixando o escopo de uma regressão simples, para uma regressão múltipla.

Outliers e Padronização dos dados:

A regressão simples, por ter somente uma variável é extremamente sensível a outliers, como no exemplo ao lado.

Sendo assim é importante antes de aplicar esse modelo fazer uma limpeza de outliers.



Outliers e Padronização dos dados:

A padronização dos dados (normalização) dos dados altera a interpretação dos coeficientes na tabela de regressão, mas muitas vezes é necessária.

Uma dica é transformar a variável x ou y aplicando log(x) ou log(y) se ela tiver uma escala muito grande.

A interpretação da tabela ficará no caso de log como na tabela ao lado.

Modelo	Interpretação
$Y = a + b*x$	1 unidade a mais em X aumenta b unidades em y
$\log(Y) = a + b*x$	1 unidade a mais em X aumenta 100*b % em y
$Y = a + b*\log(x)$	Se X aumenta em 1% , Y vai aumentar b/100 unidades
$\log(Y) = a + b*\log(x)$	Se X aumenta em 1%, Y vai aumentar em b%

Regressão Multivariada

Podemos adicionar mais variáveis explicativas para prever o valor de y. obtendo um modelo de regressão multivariada.

$$Y_i = c + b1*X1 + b2*X2 + b3*X3 + ...+ bn*Xn + e$$

Na regressão multivariada estamos diminuindo o viés de variável omitida, e adicionando no exemplo da renda variáveis explicativas como : idade, profissão entre outras variáveis para melhorar a previsão da renda.

Quantas variáveis adicionar? Podemos tentar várias, porém a medida que adicionamos mais o modelo vai ficando mais complexo o que pode gerar overfit, por isso vale monitorarmos atentamente o R-quadrado nesse caso.

Regressão Multivariada

Exemplo:

Valor da casa = c + b1*(Numero de quartos) + b2*(Tamanho em m^2) + e

No exemplo acima construímos um modelo de regressão multivariada em que queremos prever o valor da casa utilizando como variáveis o número de quartos e o tamanho em metros quadrados.

A partir do modelo vamos aplicar a minimização dos quadrados ordinários e estimar os parâmetros c, b1, b2.

Assim, tendo o número de quartos e o tamanho de um imóvel poderemos prever o seu preço, como uma calculadora!

A tabela de regressão no caso

Interpretando a tabela de regressão:

- Equação do modelo :
  y = 0.4687*x1 + 0.4836*x2 - 0.0174*x3 + 5.2058

- Nesse caso nota-se como o R-quadrado subiu, refletindo a maior complexidade do modelo

- Temos um teste F, no qual é testado se as variáveis x1, x2, x3 conjuntamente tem efeito em y

OLS Regression Results

Dep. Variable:	y	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	211.8
Date:	Fri, 13 Mar 2020	Prob (F-statistic):	6.30e-27
Time:	13:54:01	Log-Likelihood:	-34.438
No. Observations:	50	AIC:	76.88
Df Residuals:	46	BIC:	84.52
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.4687	0.026	17.751	0.000	0.416	0.522
x2	0.4836	0.104	4.659	0.000	0.275	0.693
x3	-0.0174	0.002	-7.507	0.000	-0.022	-0.013
const	5.2058	0.171	30.405	0.000	4.861	5.550

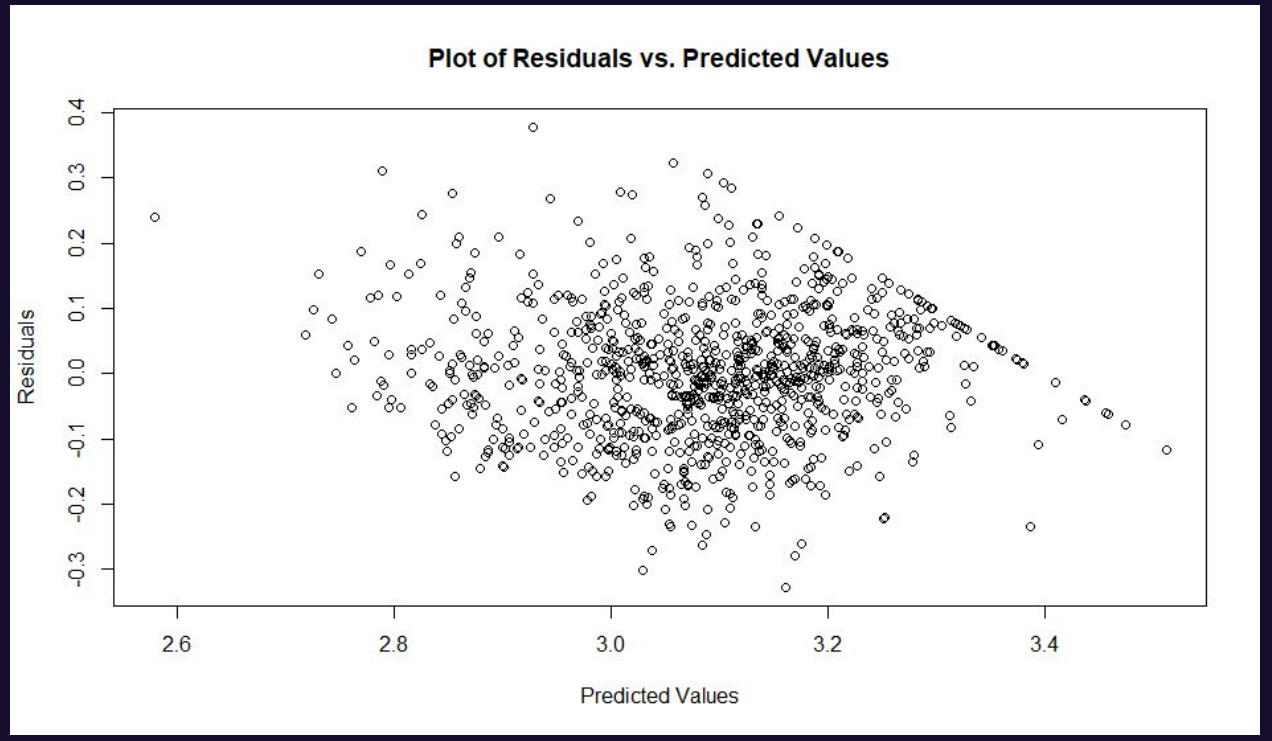
Análise de resíduos

Para garantir que um modelo de regressão foi satisfatório. Além de analisarmos o p-valor do teste t e o R quadrado devemos também olhar os gráficos de resíduo.

O Gráfico tem no eixo x o valor predito e no eixo y o resíduo.

Quando ele assume um formato de nuvem, bem disperso é um bom sinal.

Pois isso significa que o erro está "limpo" . ele nao inclui informações que deveriam ser incluídas como variável no modelo

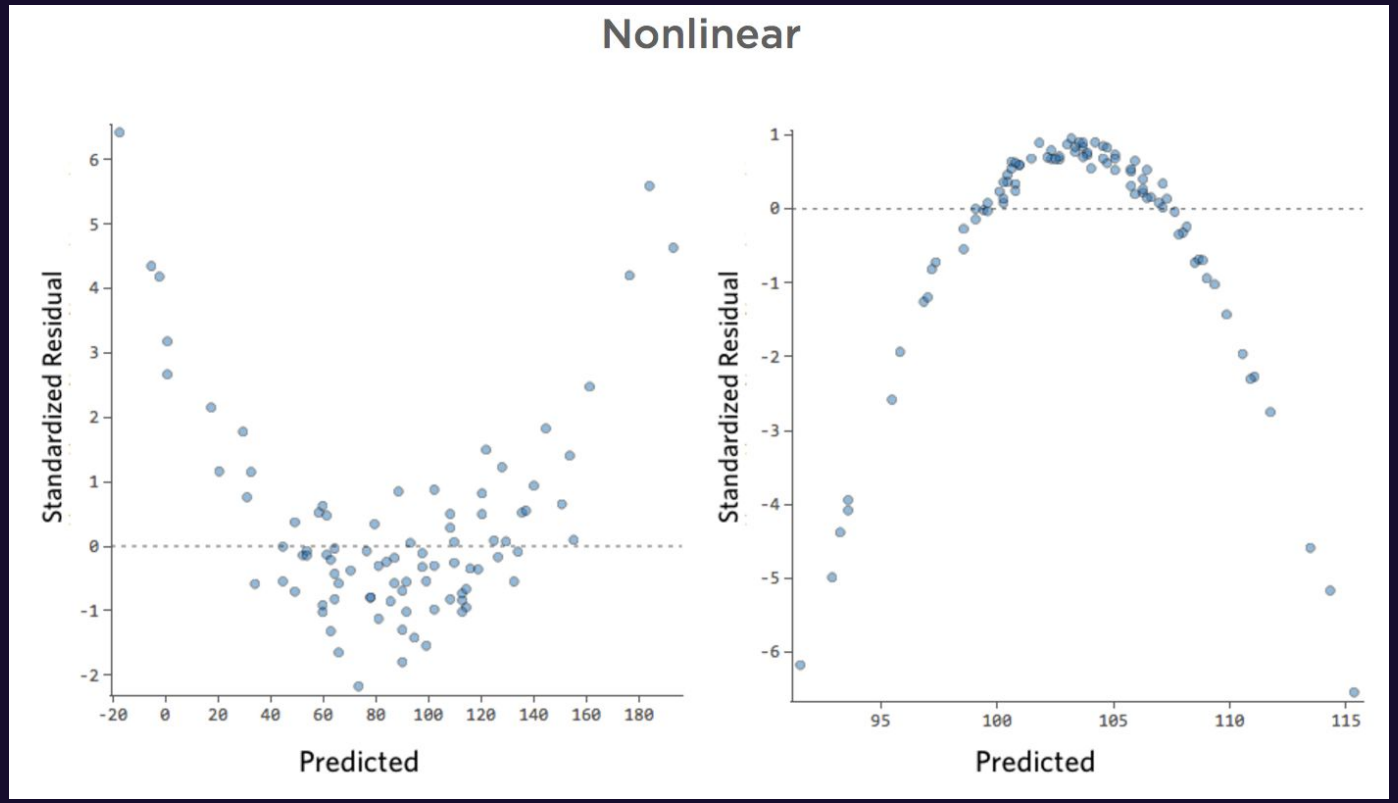


Análise de resíduos

Exemplos em que a regressão linear não cumpriu bem o seu objetivo:

No exemplo abaixo podemos ver uma relação clara entre resíduo e valor predito quadrática.

Isso significa que a regressão linear pode não ser a melhor alternativa.



Como fazemos no dia a dia?

- Rodamos um modelo de regressão quando queremos prever o valor de uma variável numérica através de outras.

- O modelo nos retorna valores de coeficientes, de modo que podemos construir uma calculadora para prever o valor a variável com uma fórmula.

- Quando uma nova informação chegar colocamos os valores na calculadora e estimamos um valor de preço ou renda ou etc.

Mas atenção o modelo só pode ser considerado válido se:

1. Os testes t tem p-valor < 0.05
2. O R-quadrado tem um bom ajuste
3. Análise do resíduo sem padrão definido.

Como fazemos no dia a dia?

O que fazemos se a regressão linear não for a melhor alternativa?

Na prática tentaremos utilizar outros modelos regressivos, por exemplo modelos de machine learning como o Random Forest Regressor, entre outras possibilidades

Assim como o modelo linear esse modelo nos trará a estimativa do variável de interesse, porém não assumindo uma relação de linearidade.

Vamos Praticar em Python!