A primeira Escola presencial gratuita de Inteligência Artificial do Brasil

UNICAMP

Itaú

accenture

FGV

everis

TATA
CONSULTANCY SERVICES

Santander

school of ai

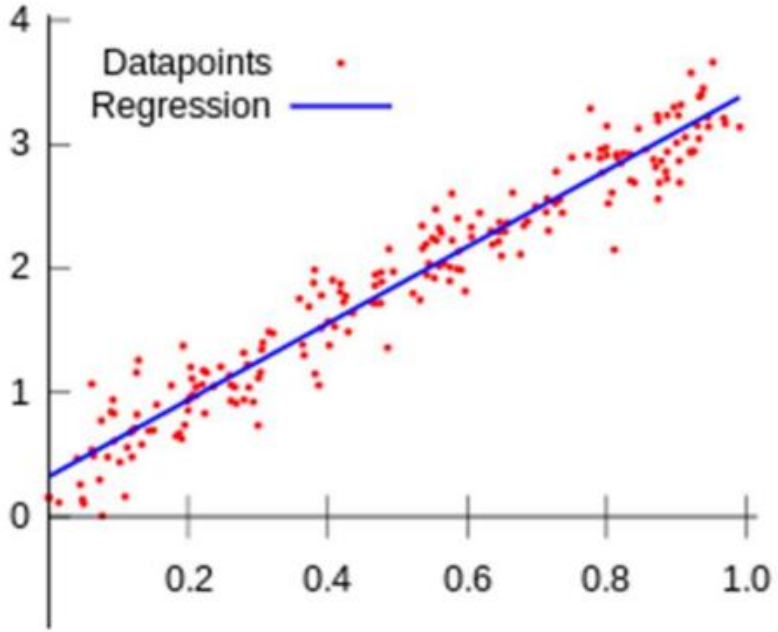
Digital
House >
Coding School

Escola
Livre de IA

Escola
Livre de IA

REVISÃO E CONCEITOS ADICIONAIS DE REGRESSÃO LINEAR

Regressão Linear



Regressão: É a função matemática que representa a sua intuição sobre como duas variáveis estão relacionadas.

Com a Curva de Regressão, podemos fazer estimativas e previsões sobre situações que não ocorreram nos casos de testes, mas que são parecidas com o que vimos na série histórica.

Francis Galton - Criou o termo Regressão enquanto estava estudando Hereditariedade.

“As espécimes que apresentam características fora do padrão tendem a ter prole com características que retornam para a média”.

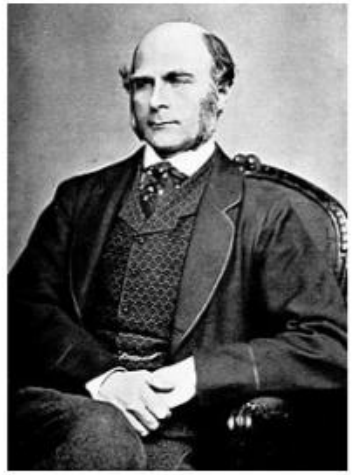
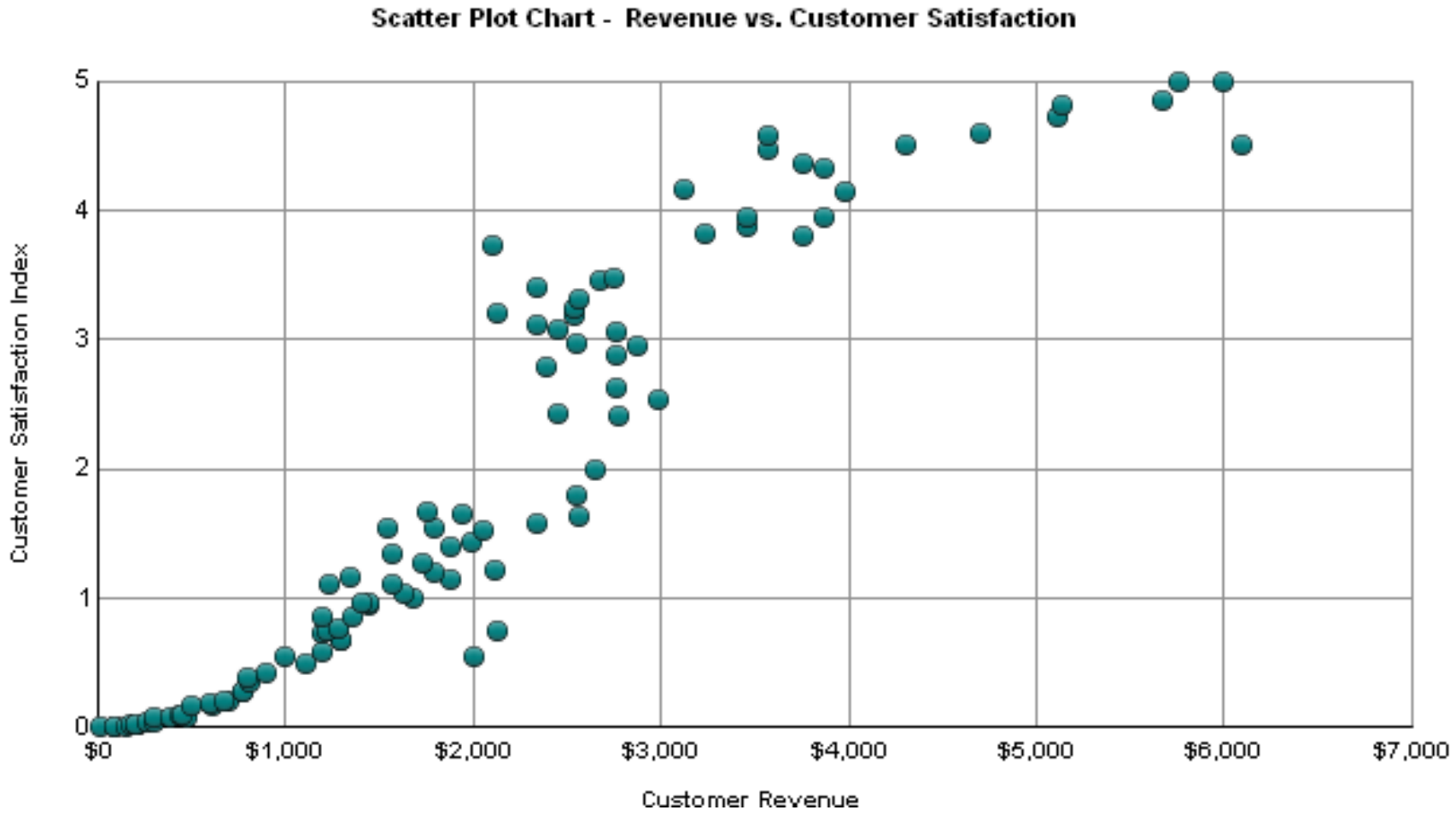


Gráfico de Dispersão

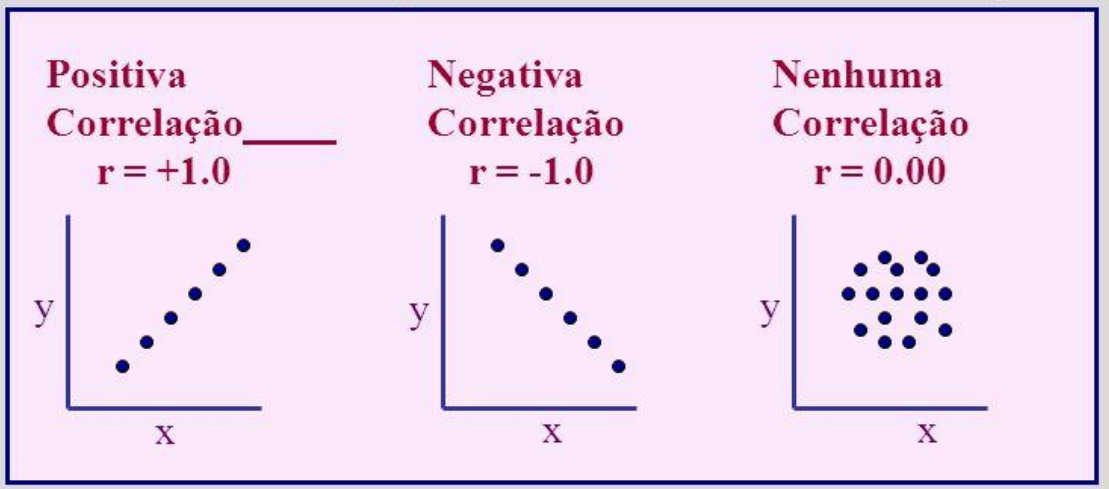


Correlação – Positiva, Negativa

Correlação

Quanto à *direção* da relação entre x e y ?

A *direção* se refere ao como os altos e baixos valores em x e y estão associados. Por exemplo:



https://slideplayer.com.br/slide/45222/

Variáveis Dependentes x Independentes

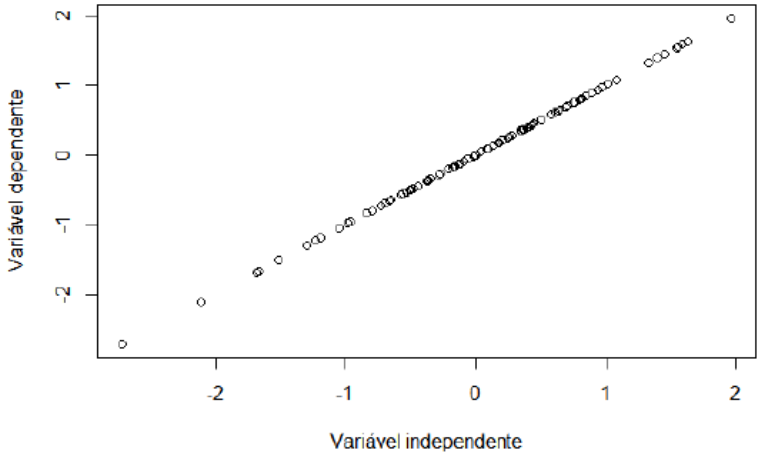


Figura 1: Correlação linear perfeita positiva.

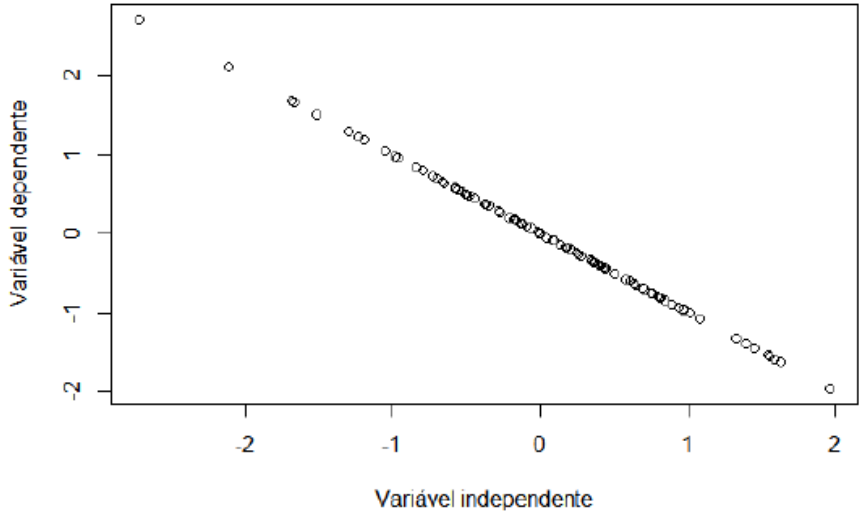


Figura 2: Correlação linear perfeita negativa.

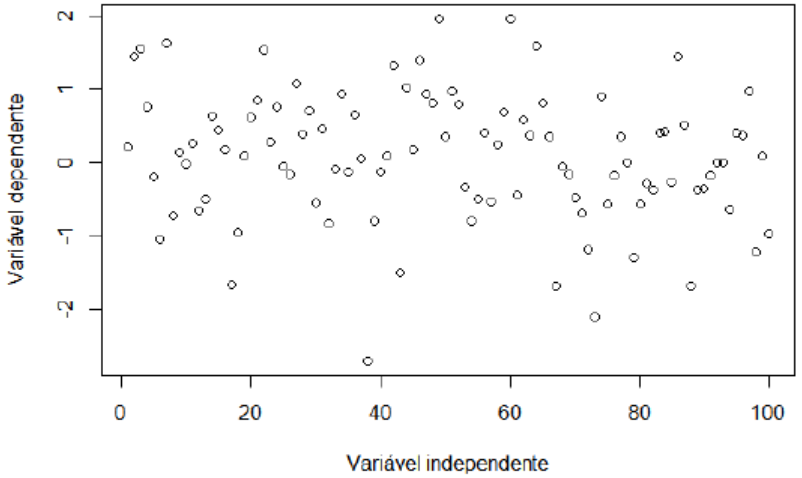


Figura 3: Não existe correlação.

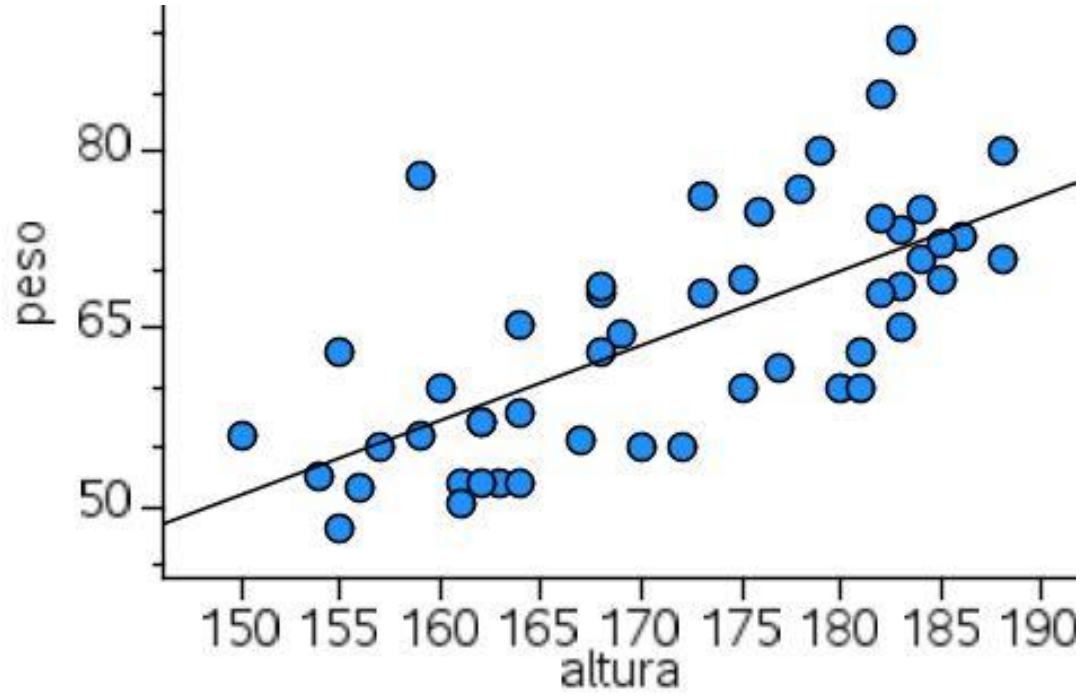
Regressão x Média

<u>Regressão Linear é a comparação entre 2 modelos:</u>

- Um modelo em que a variável independente não existe. O valor ótimo é determinado pela **média** da **variável dependente**.
- O modelo de regressão calculado.

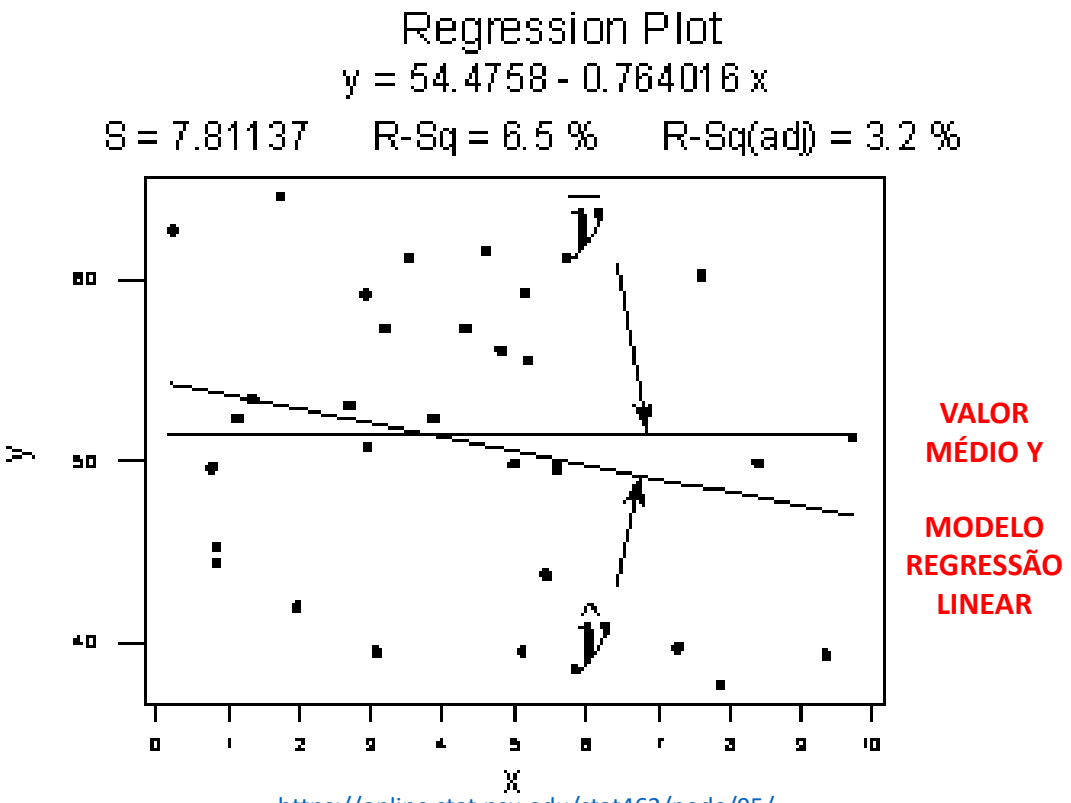
Regressão x Média

Imagine que gostaríamos de determinar o peso (var. dependente) pela altura (var. independente) da pessoa. Se não temos as informações do altura – o melhor que conseguimos é calcular o valor médio do peso.



Escola Livre de IA

Regressão x Média



VALOR MÉDIO Y

MODELO REGRESSÃO LINEAR

https://online.stat.psu.edu/stat462/node/95/

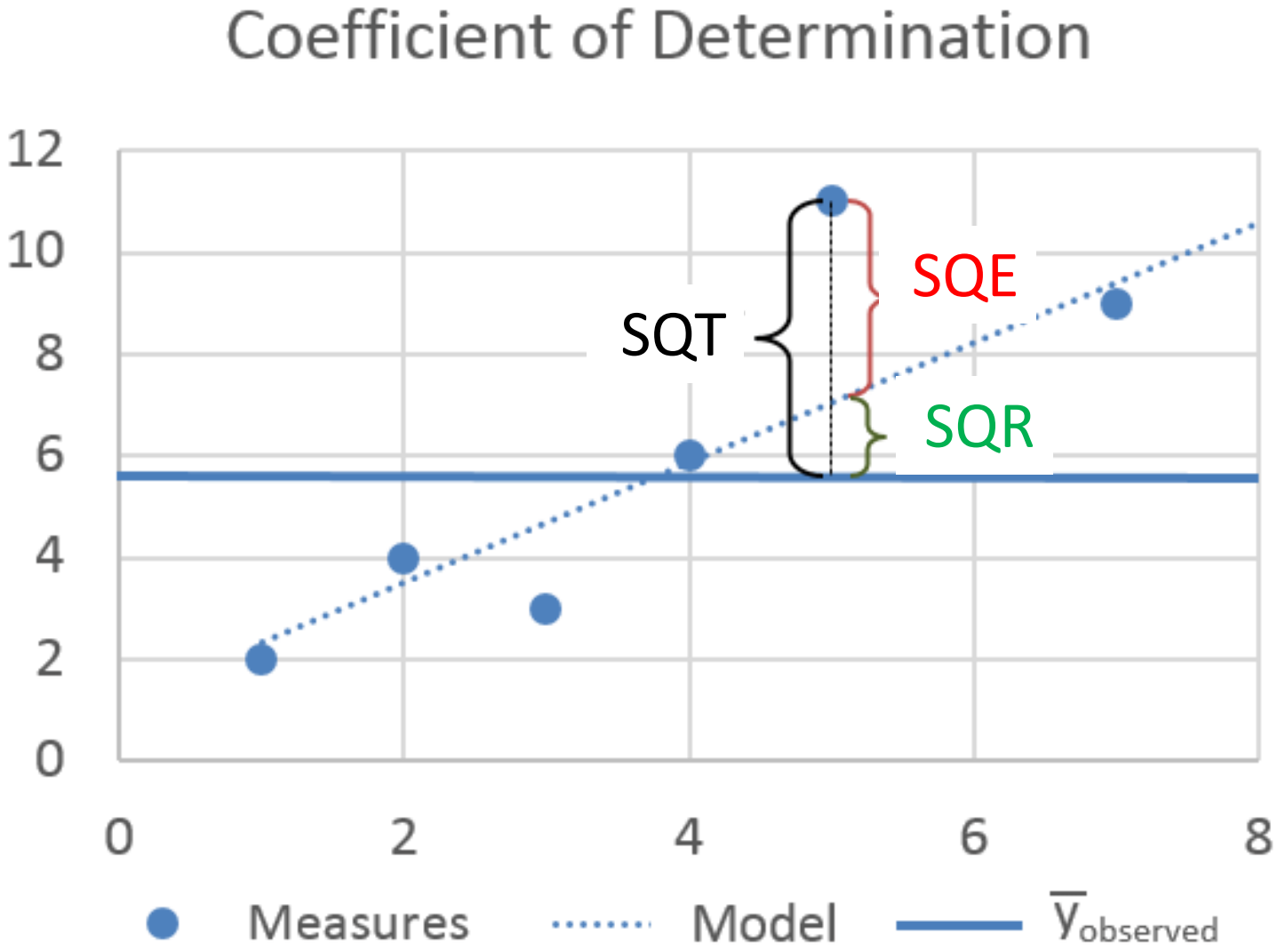
Escola Livre de IA

Somas de Quadrados

SQT = SQR + SQE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SQT	SQR	SQE
variação total	variação explicada pela equação de regressão	variação não explicada



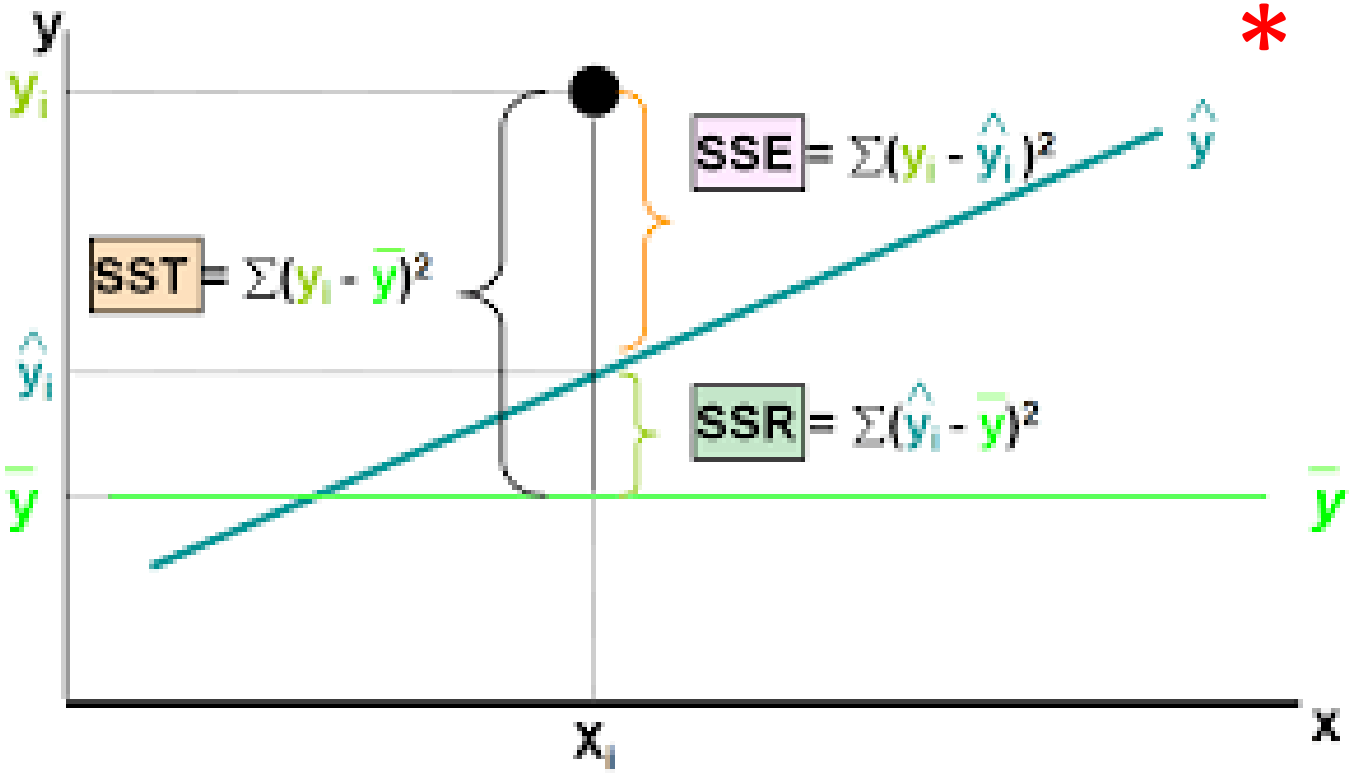
https://www.researchgate.net/figure/Visualization-of-SSE-SSR-SST_fig17_322398615

Somas de Quadrados

SQT = SQR + SQE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SQT	SQR	SQE
variação total	variação explicada pela equação de regressão	variação não explicada



*

*Observação: em inglês: SST = SSR + SSE

R^2 - Coeficiente de Determinação

Coeficiente de determinação (R^2)

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

$$\frac{SQR}{SQT}$$

$$SQT = SQR + SQE$$

Recordando Aula Regressão Linear

Outro item que é interessante analisar: R^2

- É um indicador que mede a qualidade do ajuste na Regressão Linear. Seu resultado varia de 0 a 1;

- Quanto mais próximo de 0, menos a Regressão Linear se ajustou;

- Quanto mais próximo de 1, mais a Regressão Linear se ajustou.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^{predito})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

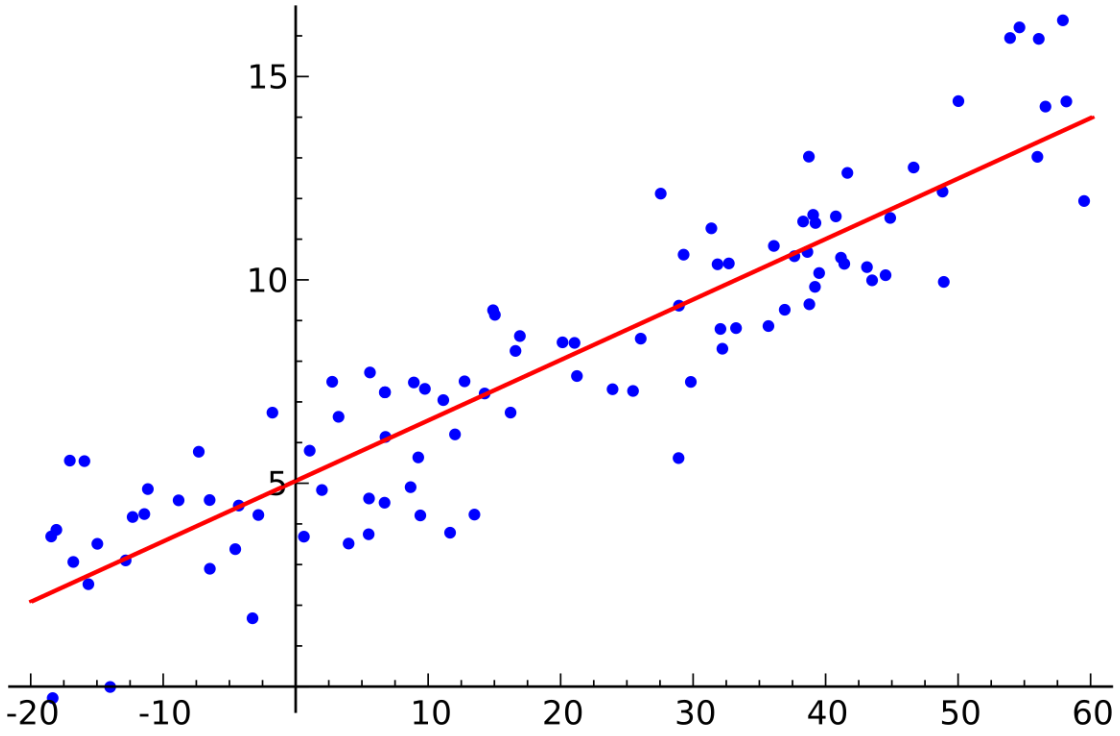
No final das contas o R^2 indica o quanto a variável dependente explica a variável independente.

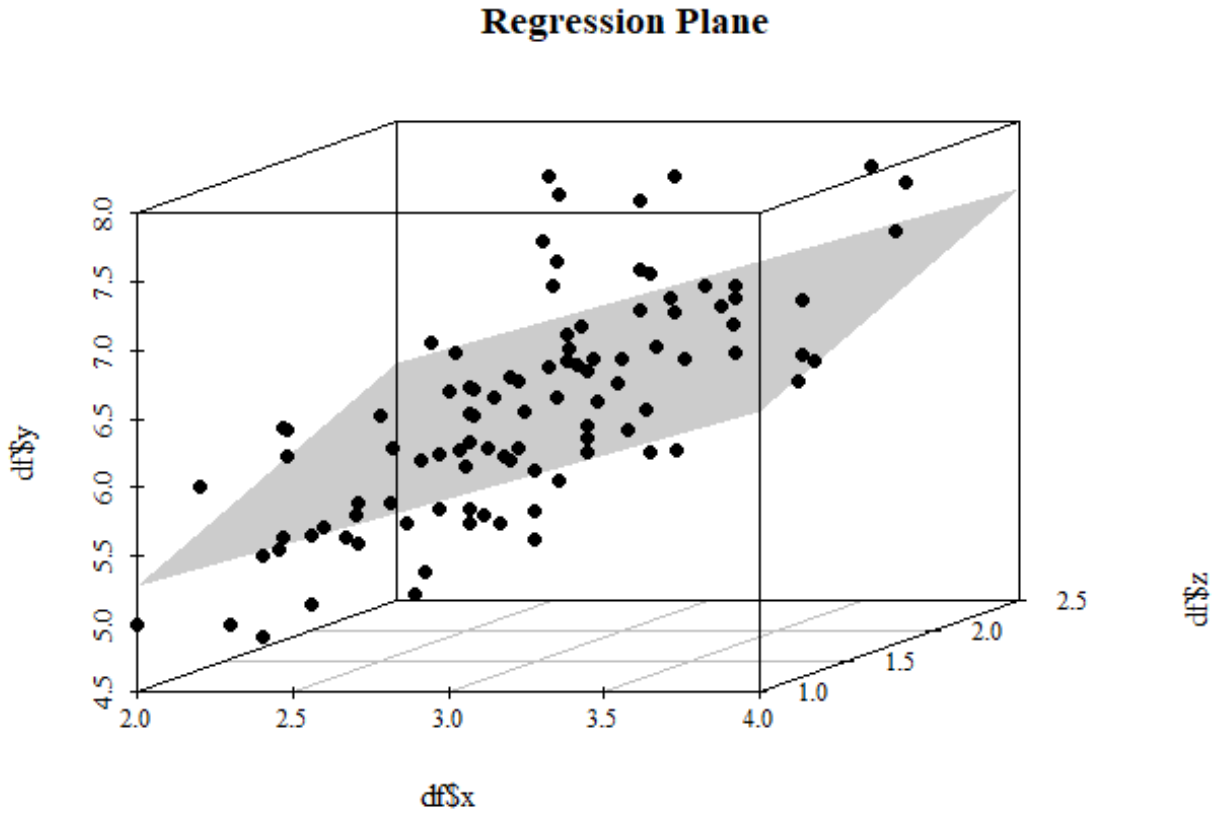
Vídeo – Coeficiente de Determinação

https://youtu.be/GE-5cLK0TM4

REGRESSÃO LINEAR MULTIPLA

Regressão Linear Múltipla

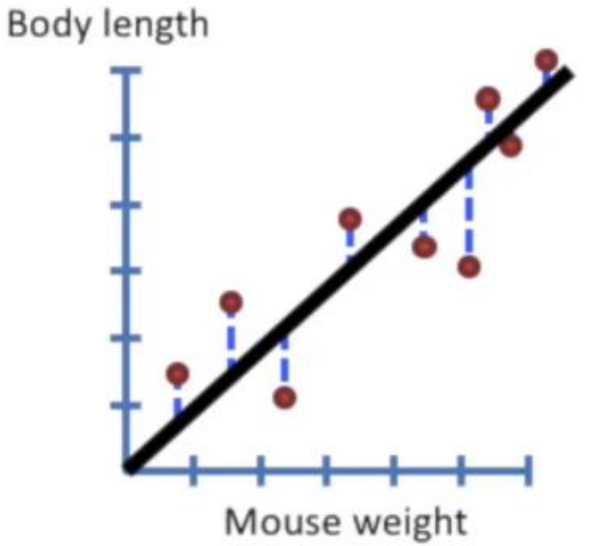




Na Regressão Linear Múltipla é uma extensão de Regressão Linear em que temos **mais do uma** variável independente.

Regressão Linear Múltipla

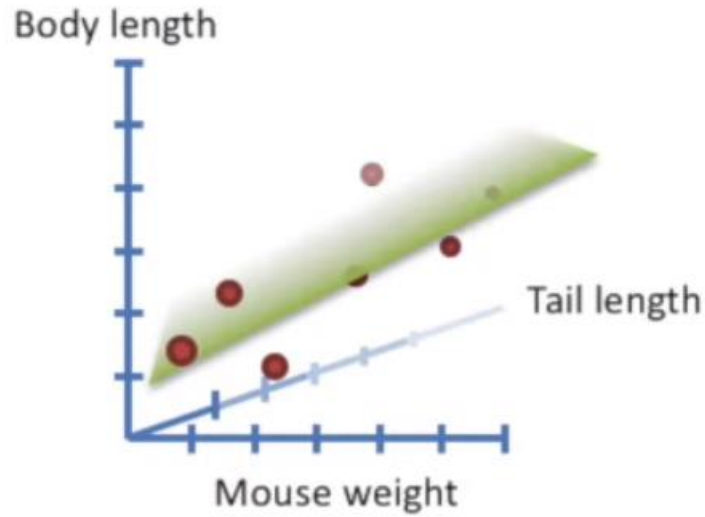
Simple regression



y = y-intercept + slope x

Na Regressão Linear Múltipla é uma extensão de Regressão Linear em que temos **mais do uma** variável independente.

Multiple regression



y = y-intercept + slope x + slope z

Escola Livre de IA

Regressão Linear Múltipla

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

Dependent variable (DV) Independent variables (IVs)

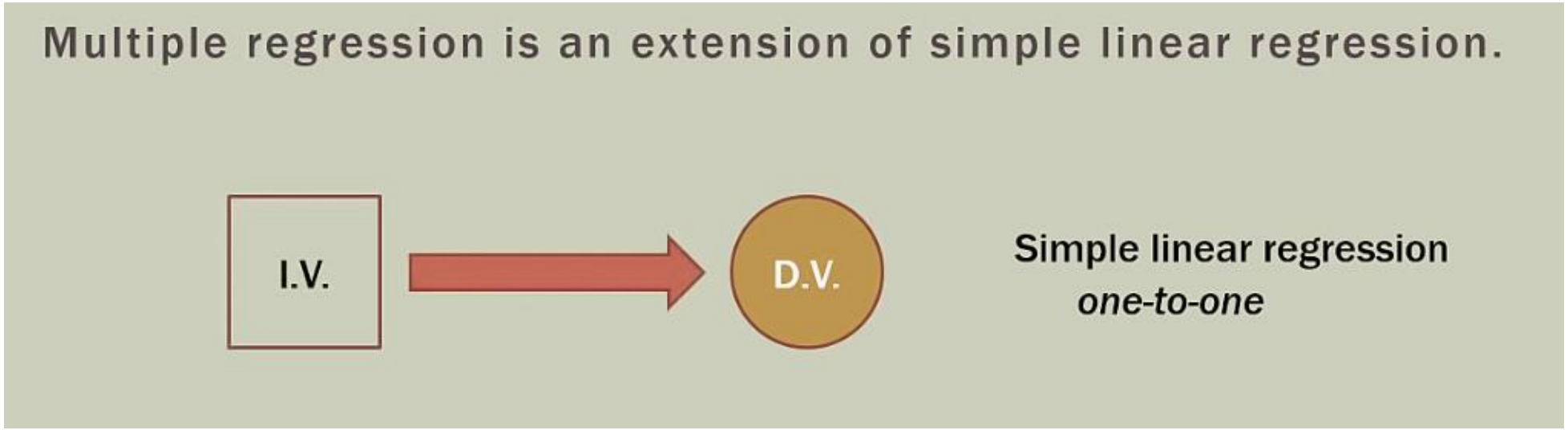
$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Vídeo - Regressão Múltipla

https://youtu.be/dQNpSa-bq4M

Escola Livre de IA

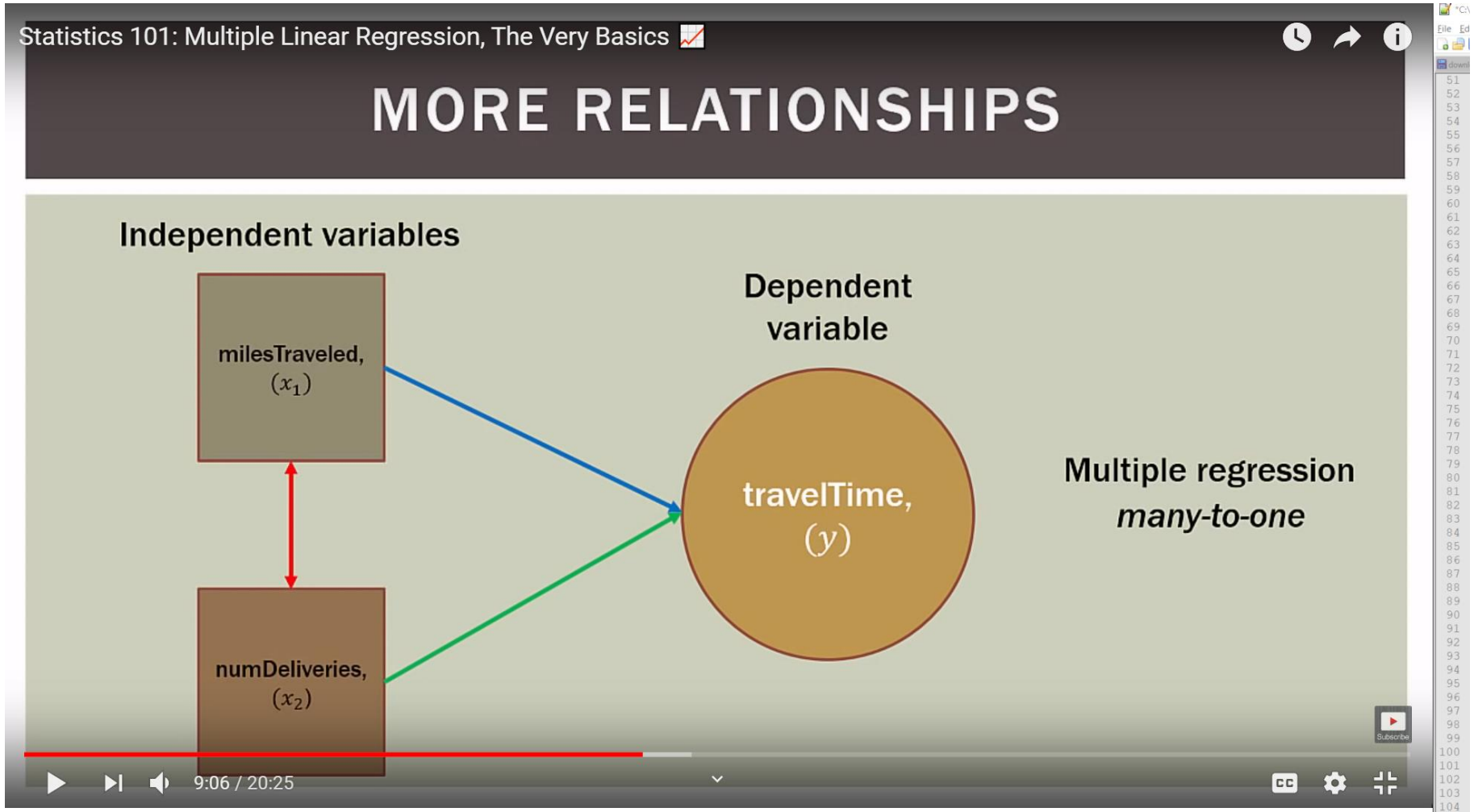
Relação entre variáveis – 1-1



https://youtu.be/dQNpSa-bq4M

Escola Livre de IA

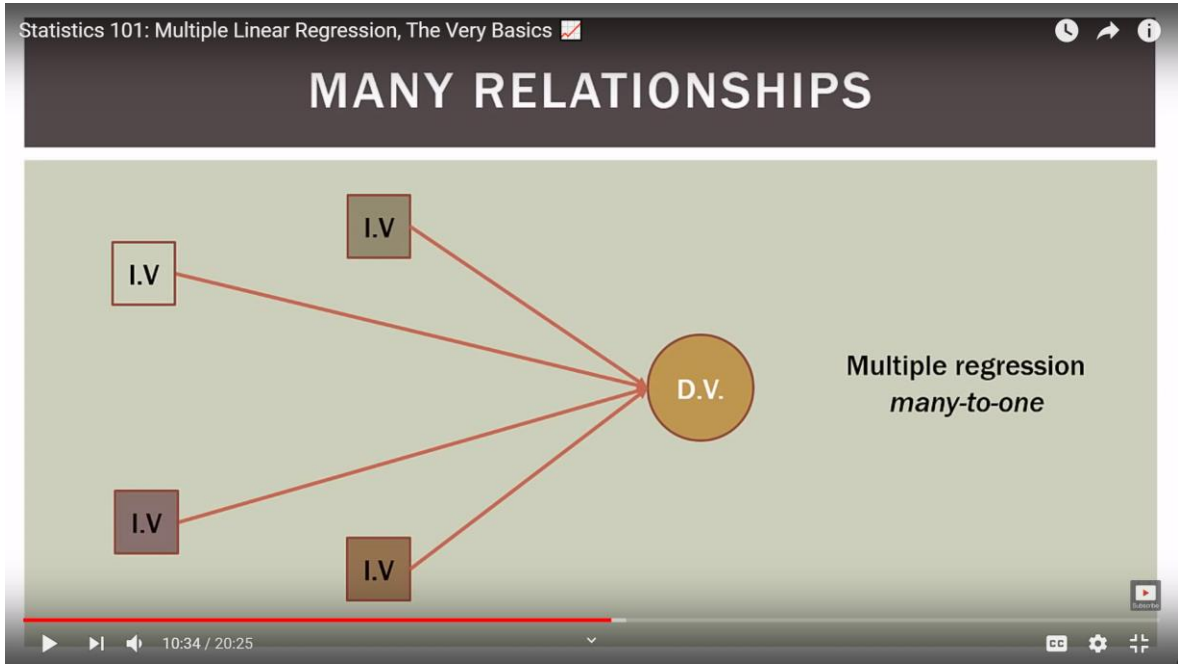
Relação entre variáveis – 1-N

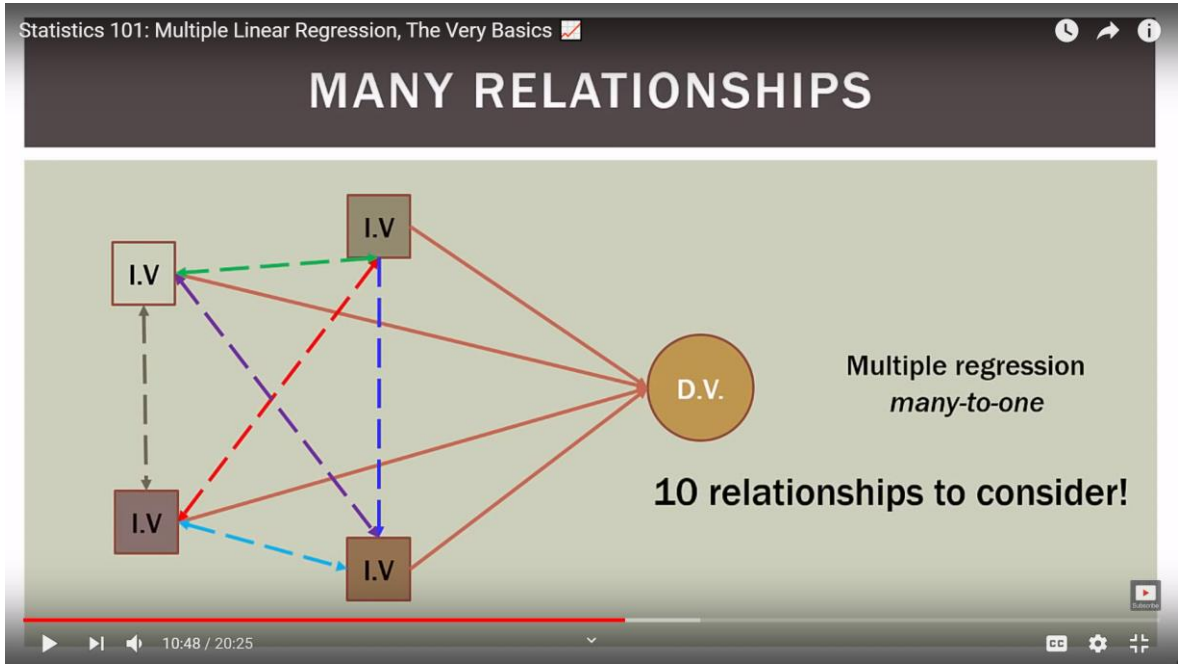


https://youtu.be/dQNpSa-bq4M

Escola Livre de IA

Relação entre variáveis – 1-N





https://youtu.be/dQNpSa-bq4M

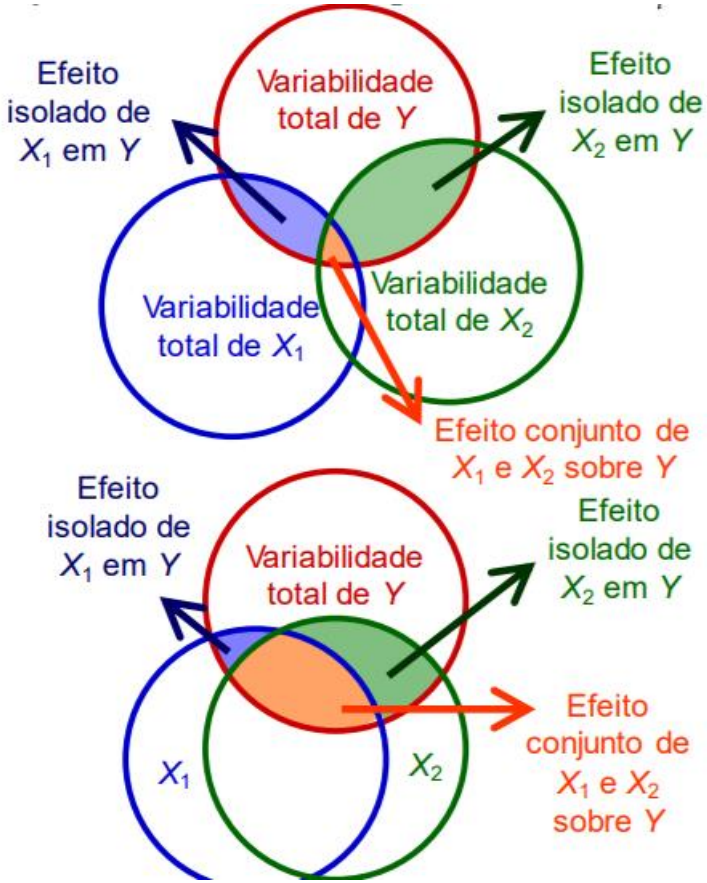
The ideal is for all of the independent variables to be correlated with the dependent variable but NOT with each other.

Escola Livre de IA

Multicolinearidade! ☹

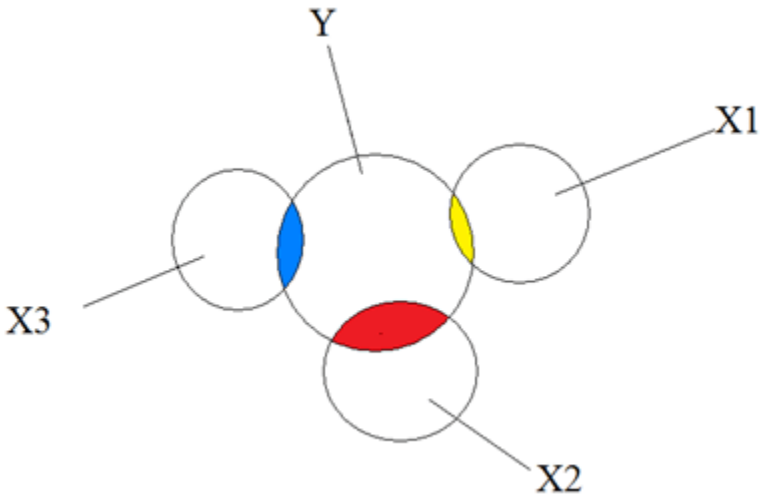
- **Multicolinearidade** consiste em um problema comum em regressões, no qual as variáveis independentes possuem **relações lineares** exatas ou aproximadamente exatas.
- O índício mais claro da existência da multicolinearidade é quando o R^2 é bastante alto, mas nenhum dos coeficientes da regressão é estatisticamente significativo segundo a estatística t convencional.
- As consequências da multicolinearidade em uma regressão são a de erros-padrão elevados no caso de multicolinearidade moderada ou severa e até mesmo a impossibilidade de qualquer estimação se a multicolinearidade for perfeita

https://pt.wikipedia.org/wiki/Multicolinearidade

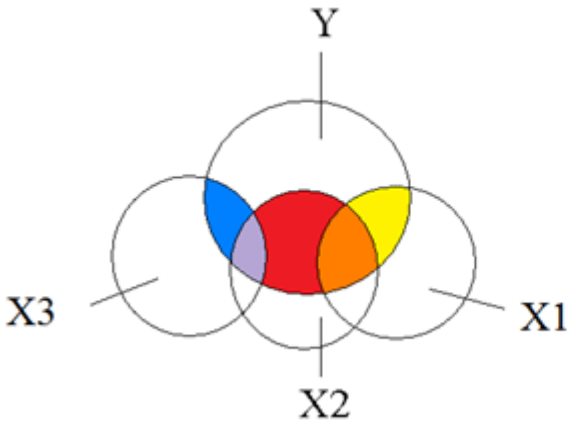


Escola Livre de IA

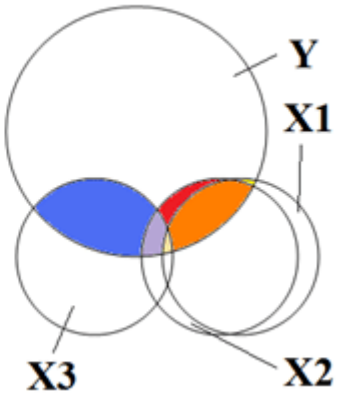
Multicolienearidade!



Sem MC



MC Moderada

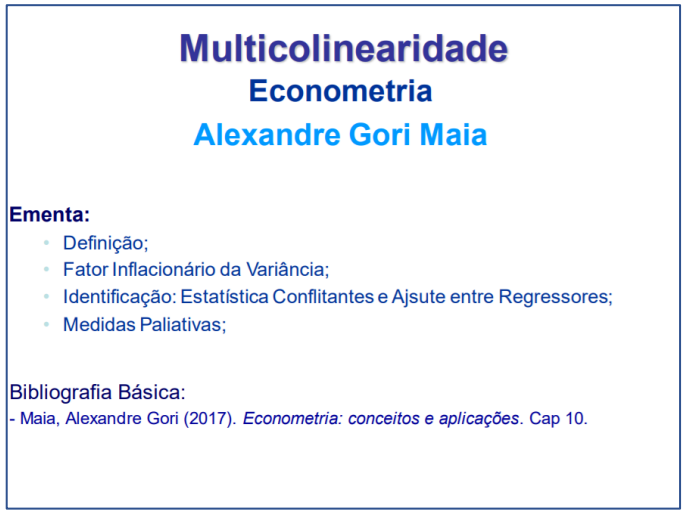


MC Extrema

https://www.theanalysisfactor.com/multicollinearity-explained-vi

Escola Livre de IA

Multicolienearidade!



Multicolinearidade

Econometria

Alexandre Gori Maia

Ementa:

- Definição;
- Fator Inflacionário da Variância;
- Identificação: Estatística Conflitantes e Ajsute entre Regressores;
- Medidas Paliativas;

Bibliografia Básica:

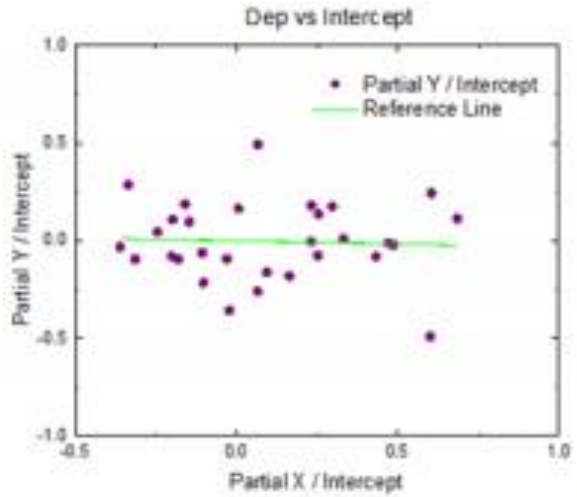
- Maia, Alexandre Gori (2017). *Econometria: conceitos e aplicações*. Cap 10.

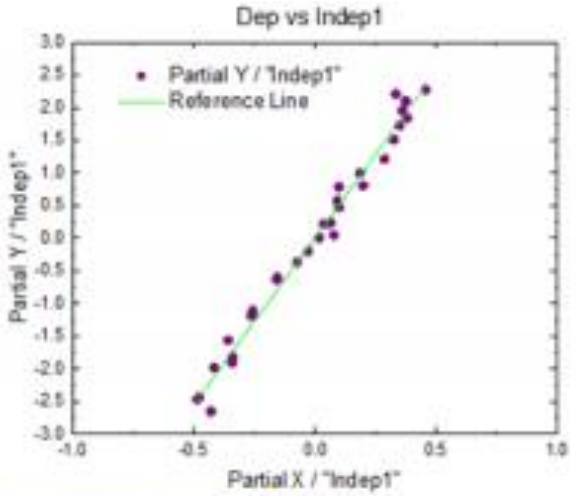
Consequências da Multicolinearidade A existência de uma colinearidade exata entre duas ou mais variáveis independentes torna impossível a obtenção dos coeficientes dos parâmetros por MQO. Por sua vez, na presença de muiltcolinearidade os estimadores de MQO continuam sendo os MELNV. O problema é que a multicolinearidade torna muitas vezes as estimativas dos coeficientes dos parâmetros insignificantes, já que cada um pressupõe, por definição, a variação em Y dada uma variação unitária em X, mantendo-se constantes as demais informações. Ou seja, se duas variáveis independentes são fortemente correlacionadas, tornár-se-á muito difícil haver variação em uma sem que haja em outra.

http://www4.eco.unicamp.br/docentes/gori/images/arquivos/EconometriaI/Ce423_Aula7_Multicolinearidade.pdf

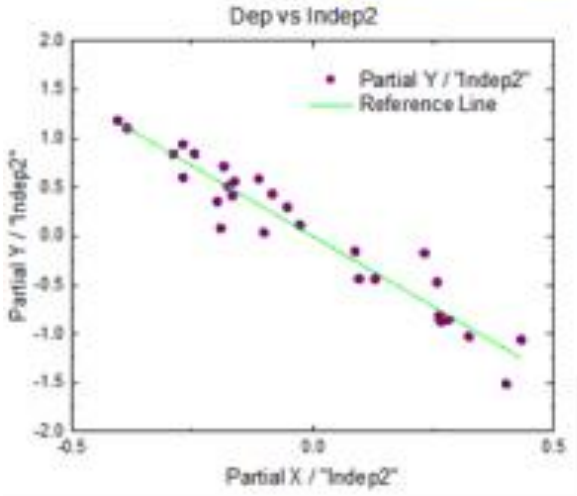
Escola Livre de IA

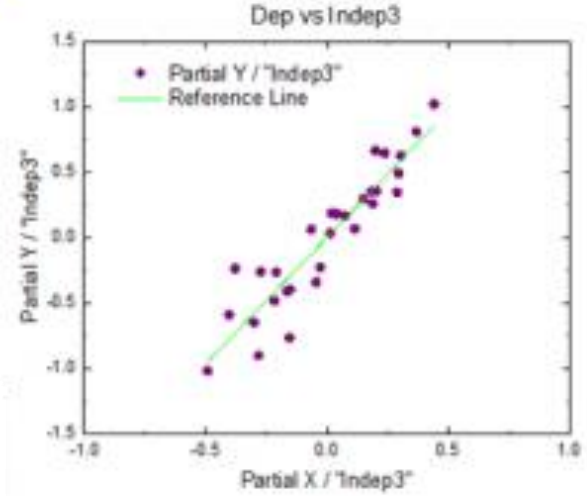
RM também é uma comparação entre modelos





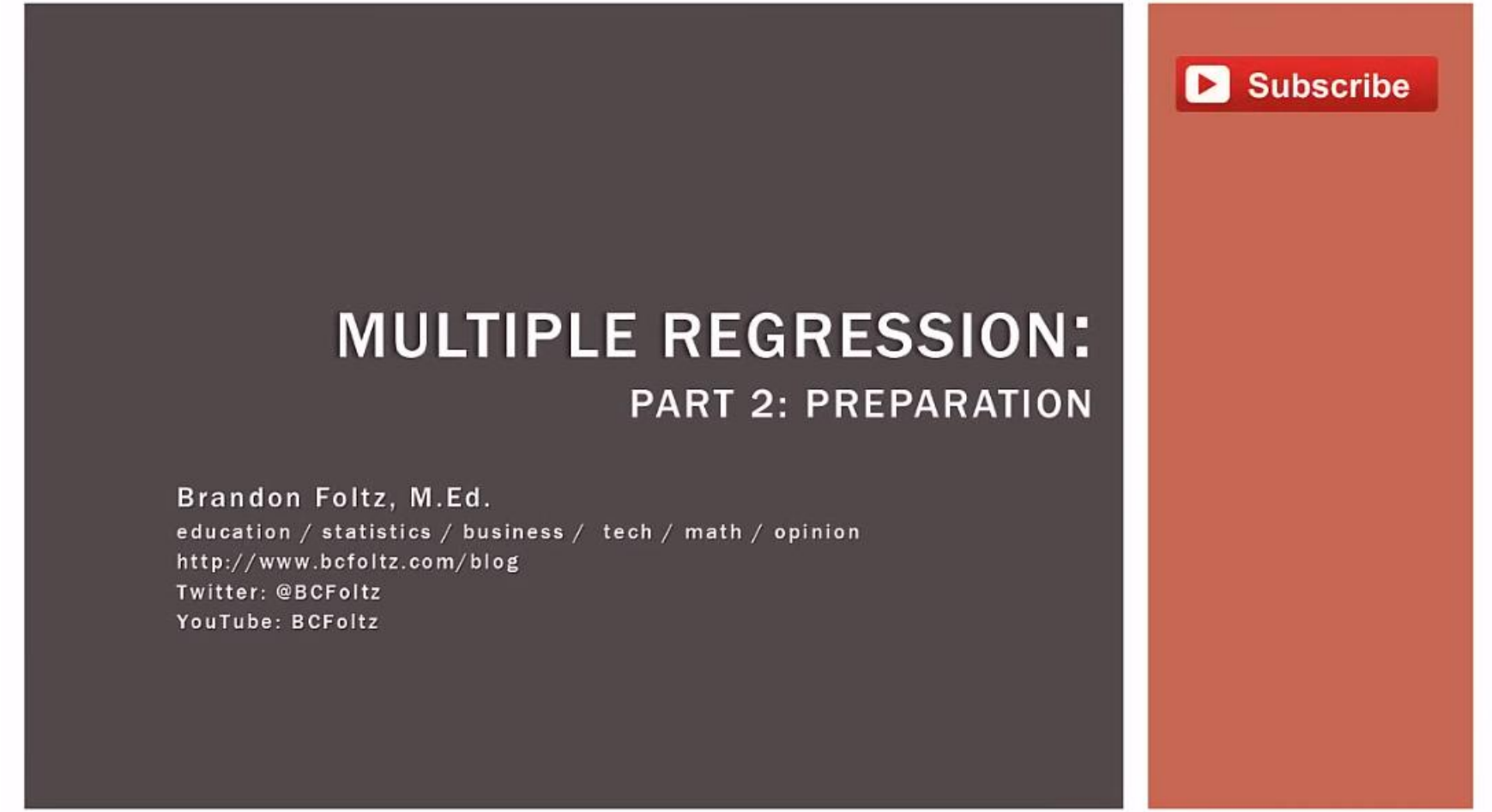
		Value	Standard Error
Dep	Intercept	-0.03365	0.1233
Dep	"Indep1"	5.10482	0.13219
Dep	"Indep2"	-2.88676	0.15718
Dep	"Indep3"	1.93475	0.15656





Escola Livre de IA

Vídeo – Preparação para RM



https://youtu.be/2I_AYIECCOQ

Escola Livre de IA

Problema Apresentado

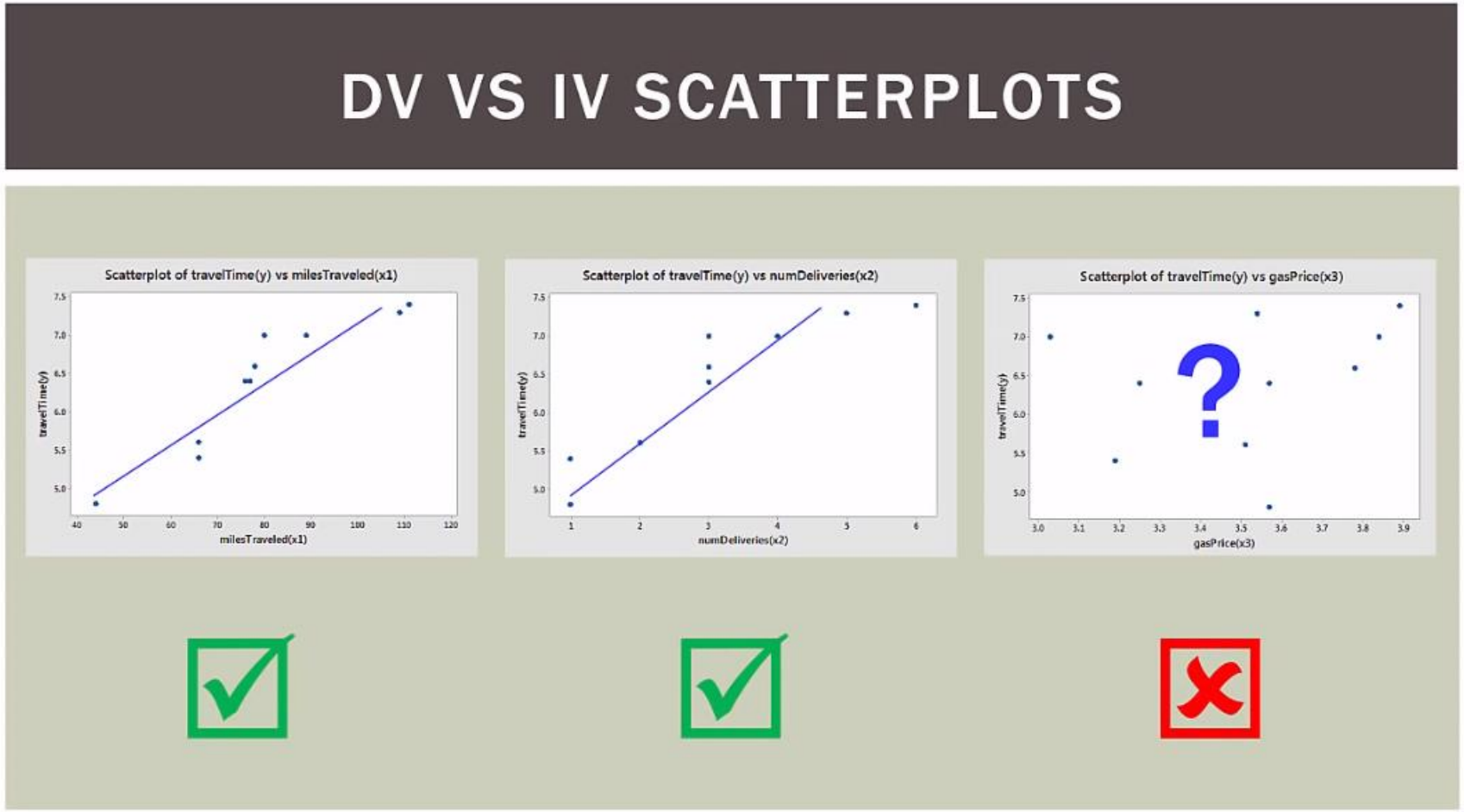
RDS DATA AND VARIABLE NAMING

To conduct your analysis you take a random sample of 10 past trips and record four pieces of information for each trip: 1) total miles traveled, 2) number of deliveries, 3) the daily gas price, and 4) total travel time in hours.

milesTraveled,(x_1)	numDeliveries,(x_2)	gasPrice,(x3)	travelTime(hrs),(y)
89	4	3.84	7
66	1	3.19	5.4
78	3	3.78	6.6
111	6	3.89	7.4
44	1	3.57	4.8
77	3	3.57	6.4
80	3	3.03	7
66	2	3.51	5.6
109	5	3.54	7.3
76	3	3.25	6.4

Queremos determinar o **tempo** (em horas – variável dependente) em função da **distância percorrida**, **quantidade de entregas** e **preço do combustível** (variáveis independentes)

Comparação da Var. Dependente com as Var. Ind.



Nós desejamos que as variáveis Independentes sejam correlacionadas com a Dependente.

Escola Livre de IA

Comparação entre Var. Independentes

Por outro lado não desejamos que as variáveis sejam correlacionadas entre si – o que implicaria em Multicolinearidade.

Escola Livre de IA

Comparação entre modelos

Regressões Lineares

Regressões Múltiplas

MODEL OPTIONS SUMMARY

F	$p-$value	S	$R^2(adj)$	$R^2(pred)$	x_1	x_2	x_3	VIF
49.77	< 0.001	0.34230	84.42%	79.07%	X			1.00
41.96	< 0.001	0.36809	81.99%	70.27%		X		1.00
0.62	0.455	0.88640	0.00%	0.00%			X	1.00
23.72	0.001	0.35264	83.47%	59.95%	X	X		11.59
22.63	0.001	0.35988	82.78%	68.11%	X		X	1.14
27.63	< 0.001	0.32970	85.55%	71.76%		X	X	1.33

"One way to measure multicollinearity is the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be 1. A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity."

Source: http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis

Comparação entre modelos

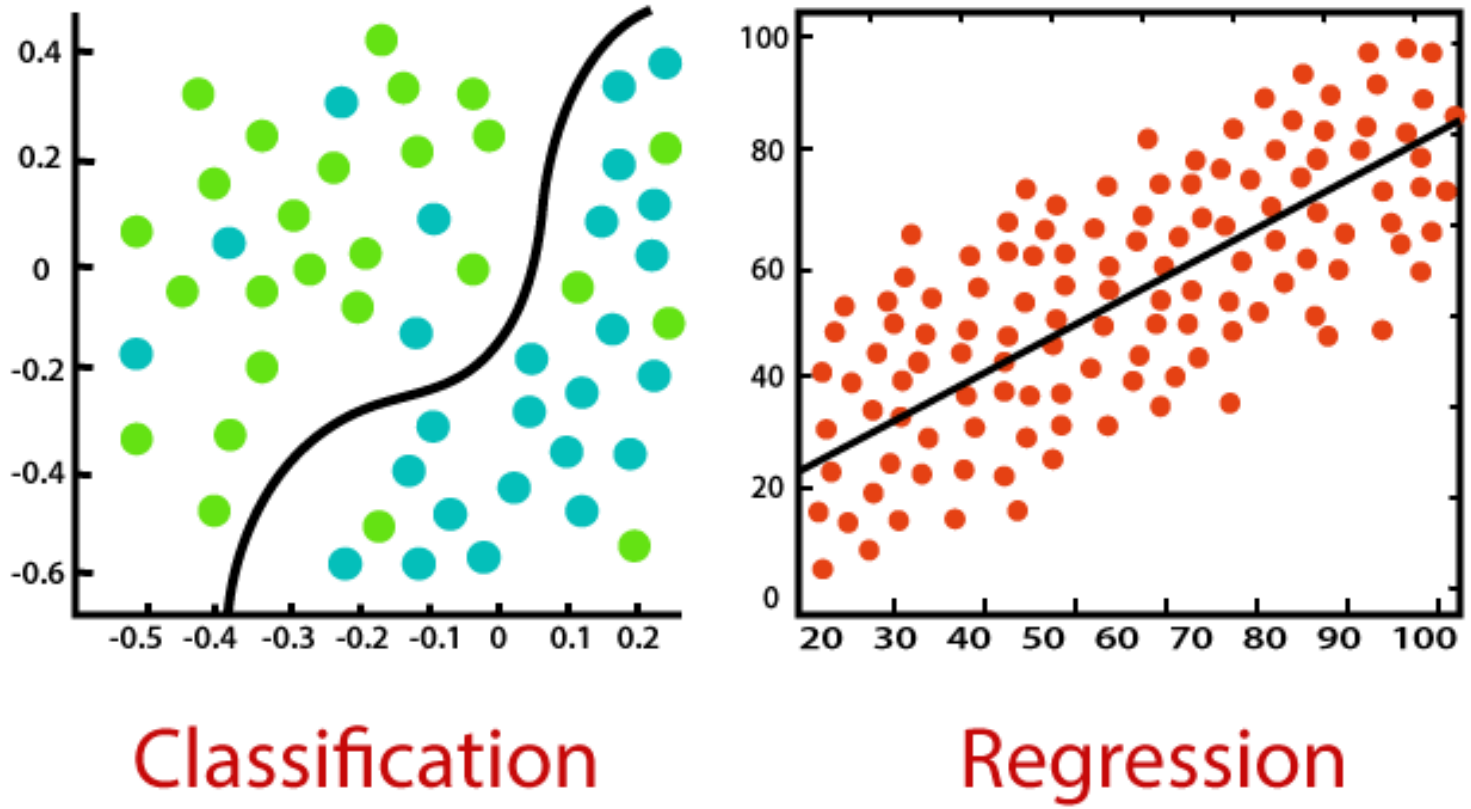
Best Subsets Regression: travelTime(y versus milesTravele, numDeliverie, ...

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	x1	x2	x3
1	86.2	84.4	79.1	1.9	0.34231	X		
1	84.0	82.0	70.3	3.1	0.36809		X	
2	88.8	85.5	71.8	2.4	0.32970		X	X
2	87.1	83.5	59.9	3.3	0.35264	X	X	
3	89.5	84.2	57.5	4.0	0.34469	X	X	X

1. Look at R-Sq(adj). Which are the highest values?
2. Look at R-Sq(pred). Which are the highest values?
3. Examine the difference between R-Sq(adj) and R-Sq(pred). A large drop-off indicates overfitting; too many variables in the model.
4. Look at Mallows C_p. Look for one that is low and approximately equals the number of predictors plus the constant (1).
5. Using all of the above, choose the best model.

REGRESSÃO LOGÍSTICA

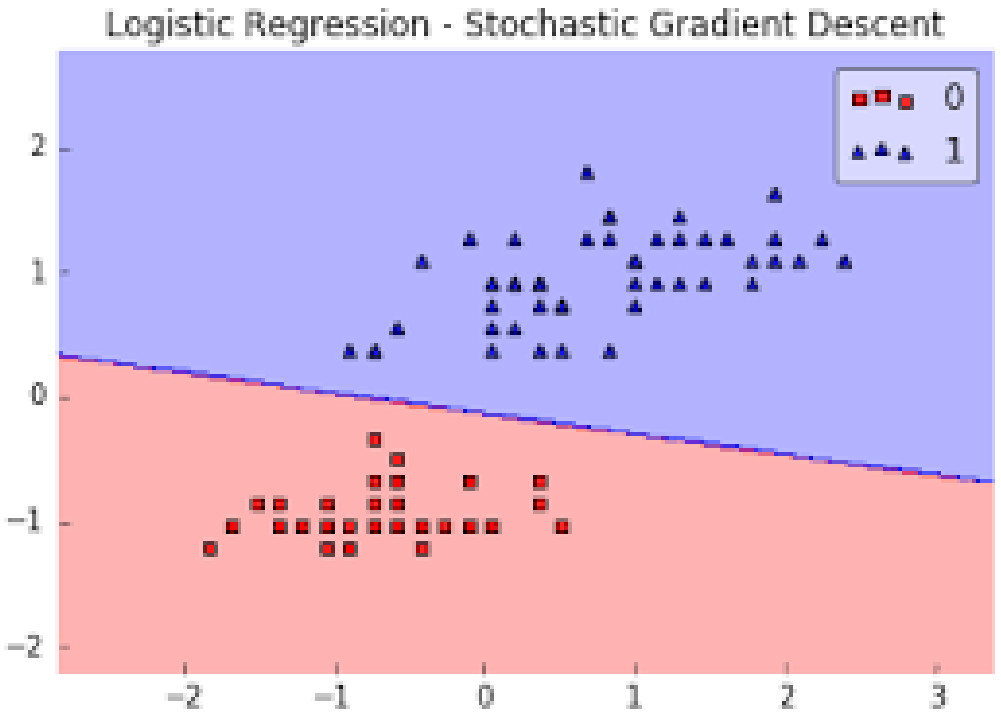
Regressão Logística - Classificação



Regressão Logística é utilizada para **CLASSIFICAR** valores **Discretos** – em oposição a **Regressão Linear** e **Regressão Múltipla** que predizem valores **contínuos**.

Escola Livre de IA

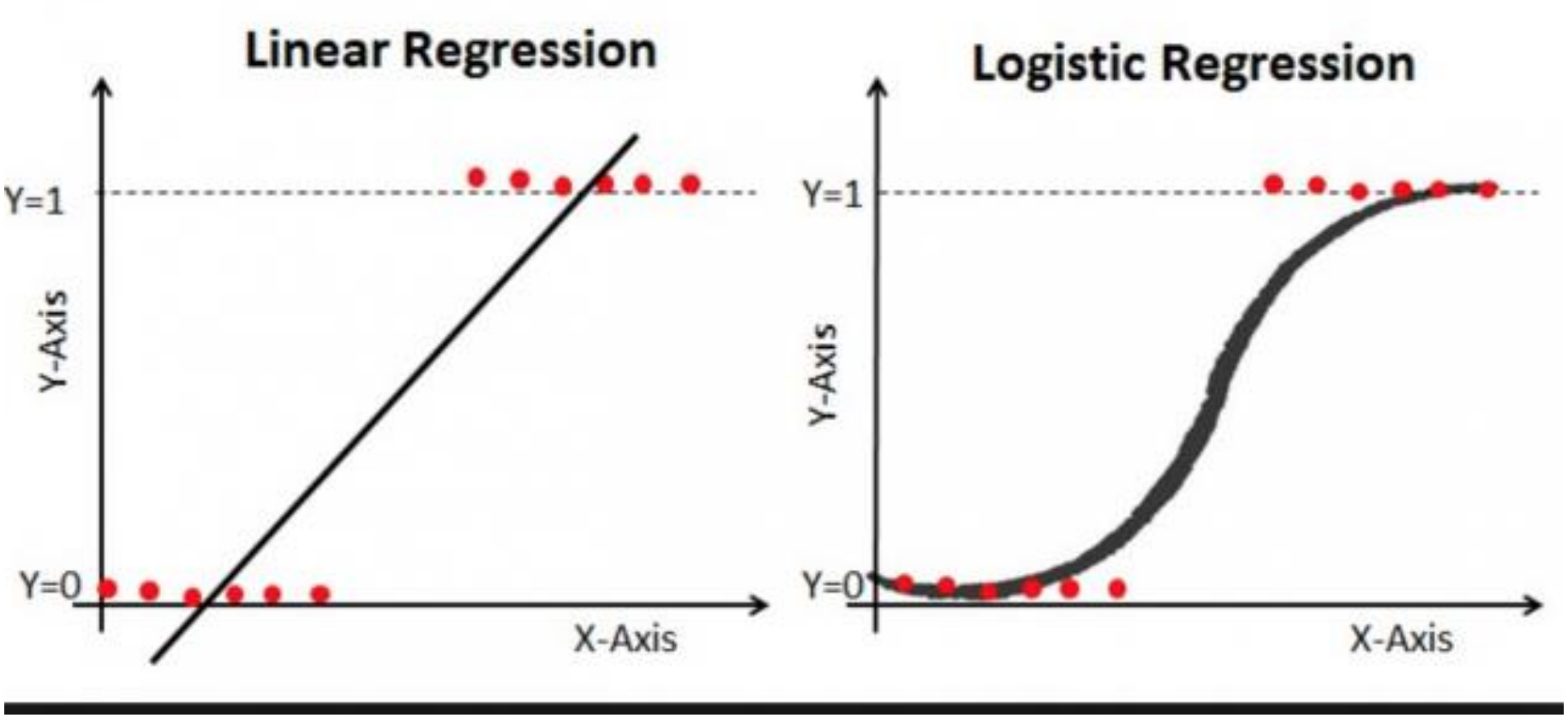
Regressão Logística - Classificação



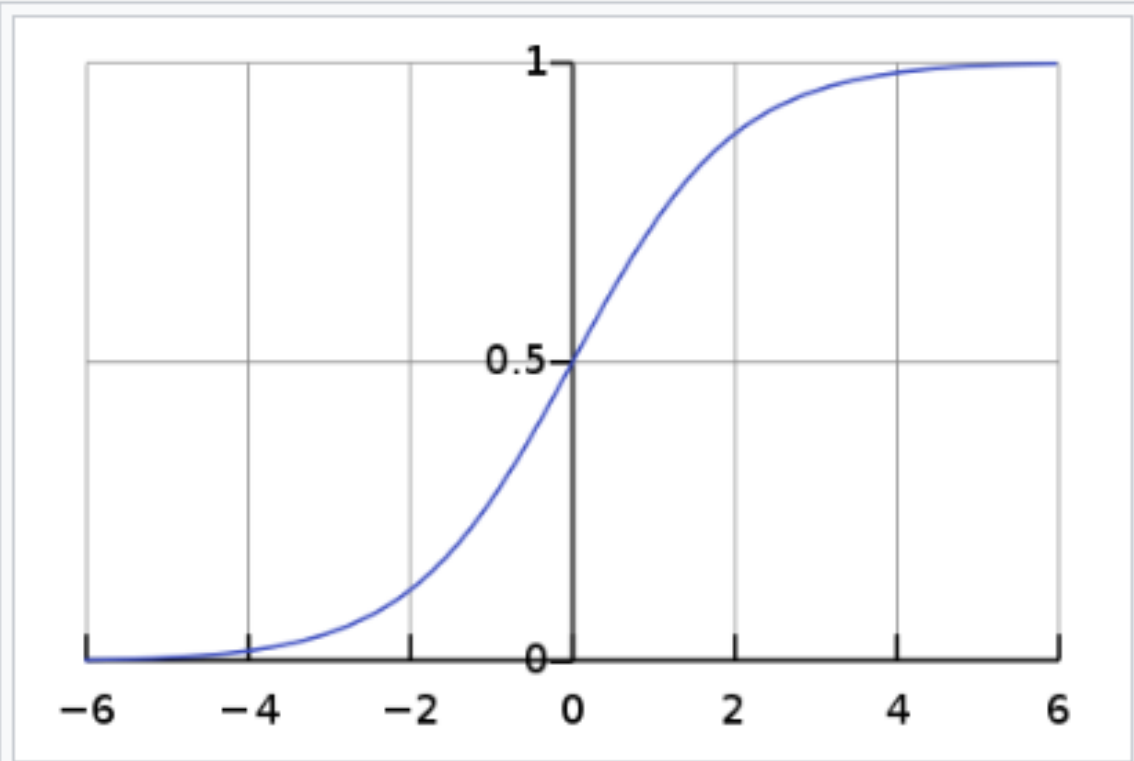
Regressão Logística é utilizada para CLASSIFICAR valores Discretos – em oposição a Regressão Linear e Regressão Múltipla que predizem valores contínuos.

Escola Livre de IA

Por que RL não funciona?



Função Logística Sigmoide



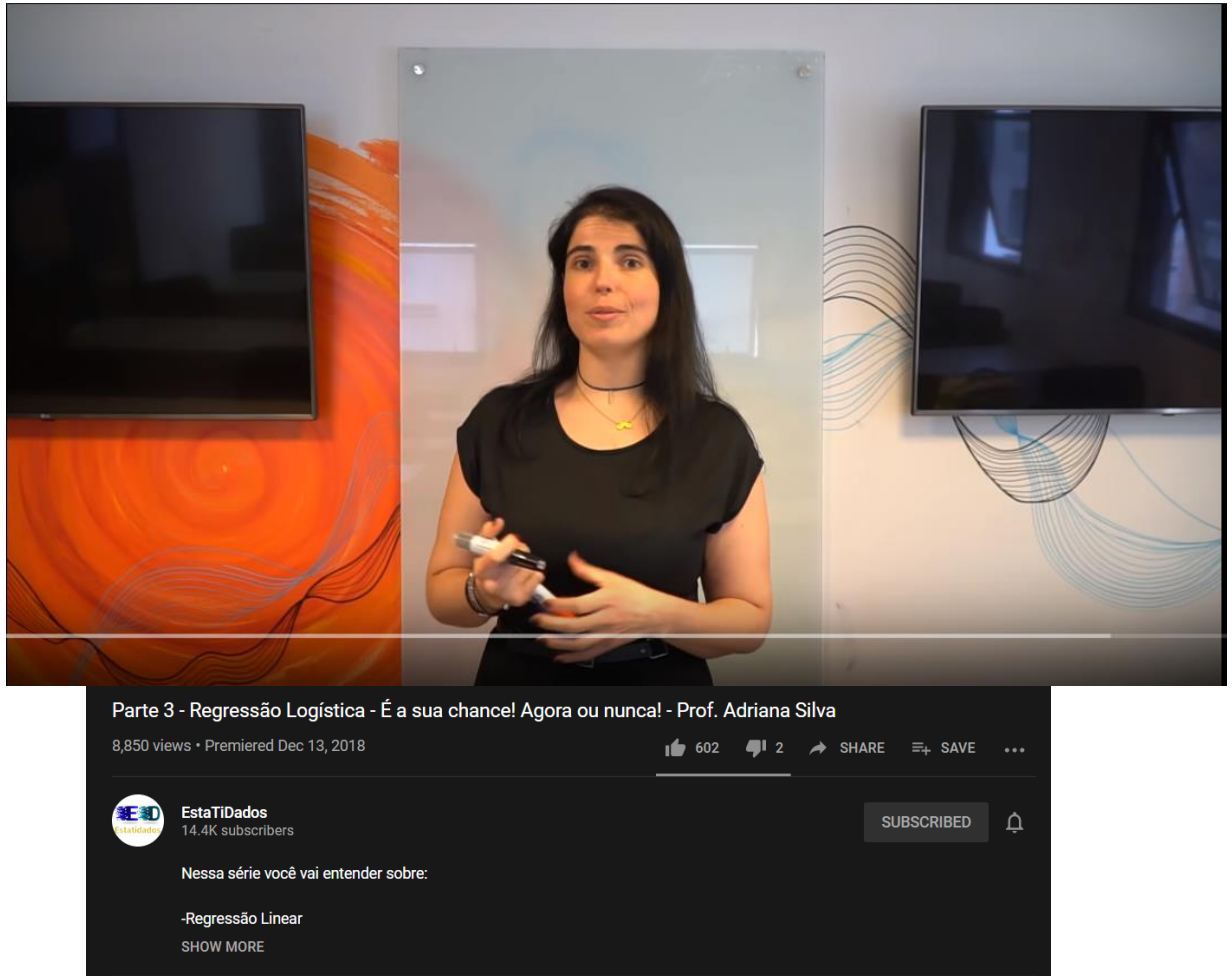
Função logística sigmoide padrão

https://pt.wikipedia.org/wiki/Fun%C3%A7%C3%A3o_log%C3%ADstica

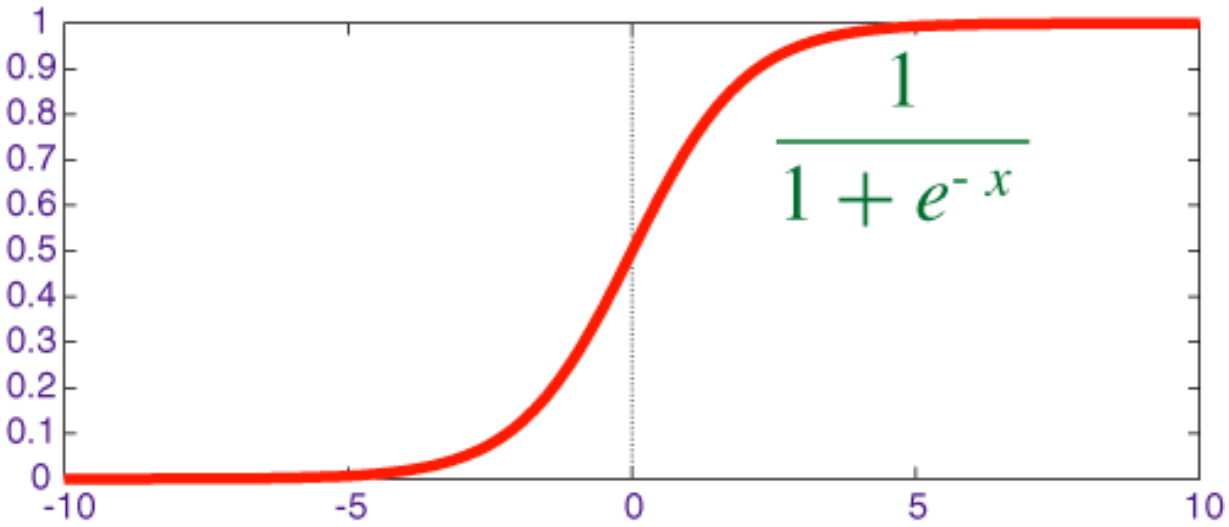
A ideia principal é que iremos o considerar o valor resultante como uma PROBABILIDADE daquilo que queremos classificar.

Escola Livre de IA

Como chegamos à Sigmoide?



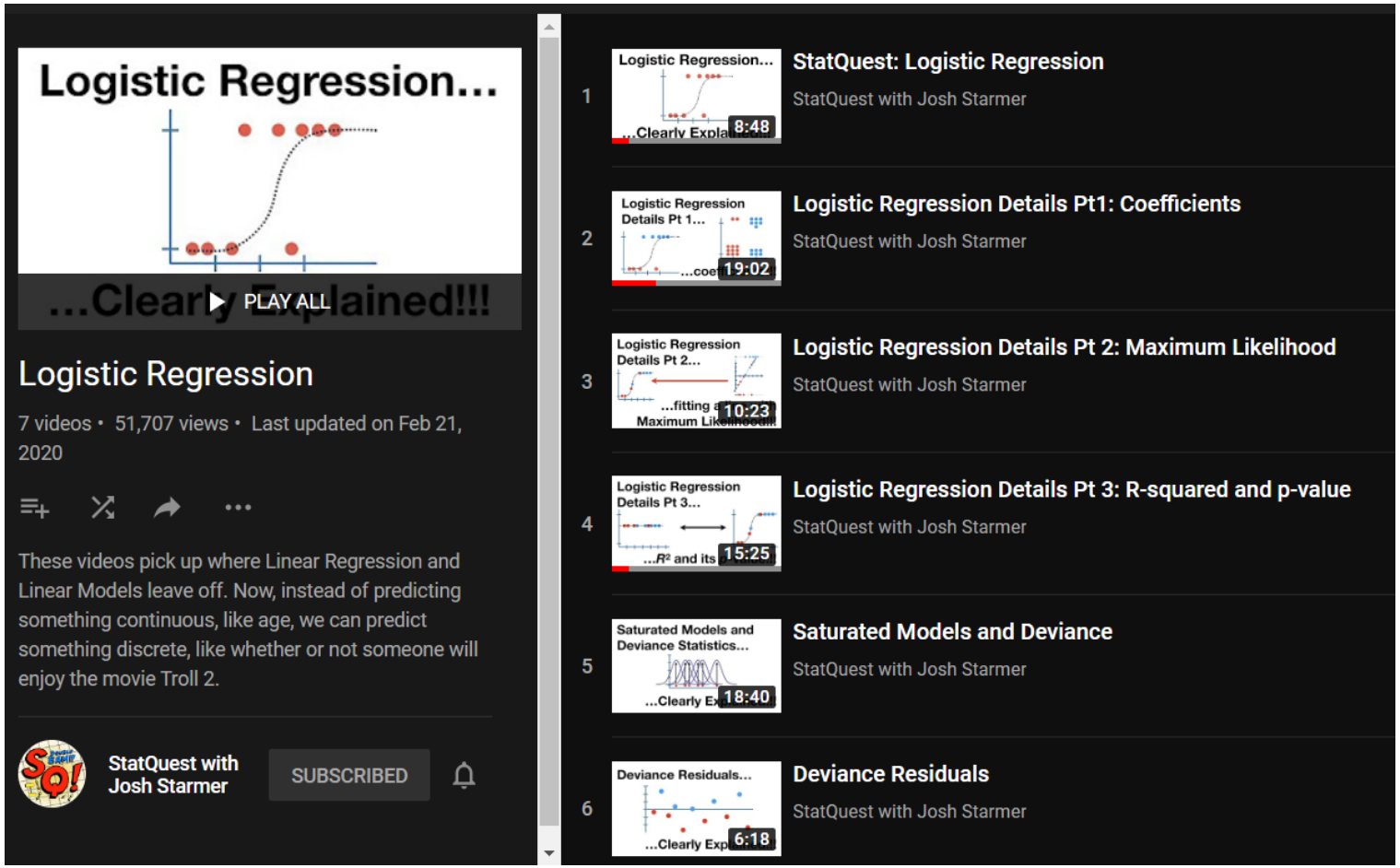
https://youtu.be/dcsZsA_wipE



https://www.i2tutorials.com/top-machine-learning-interview-questions-and-answers/what-is-sigmoid-function-and-explain-in-detail/

Escola Livre de IA

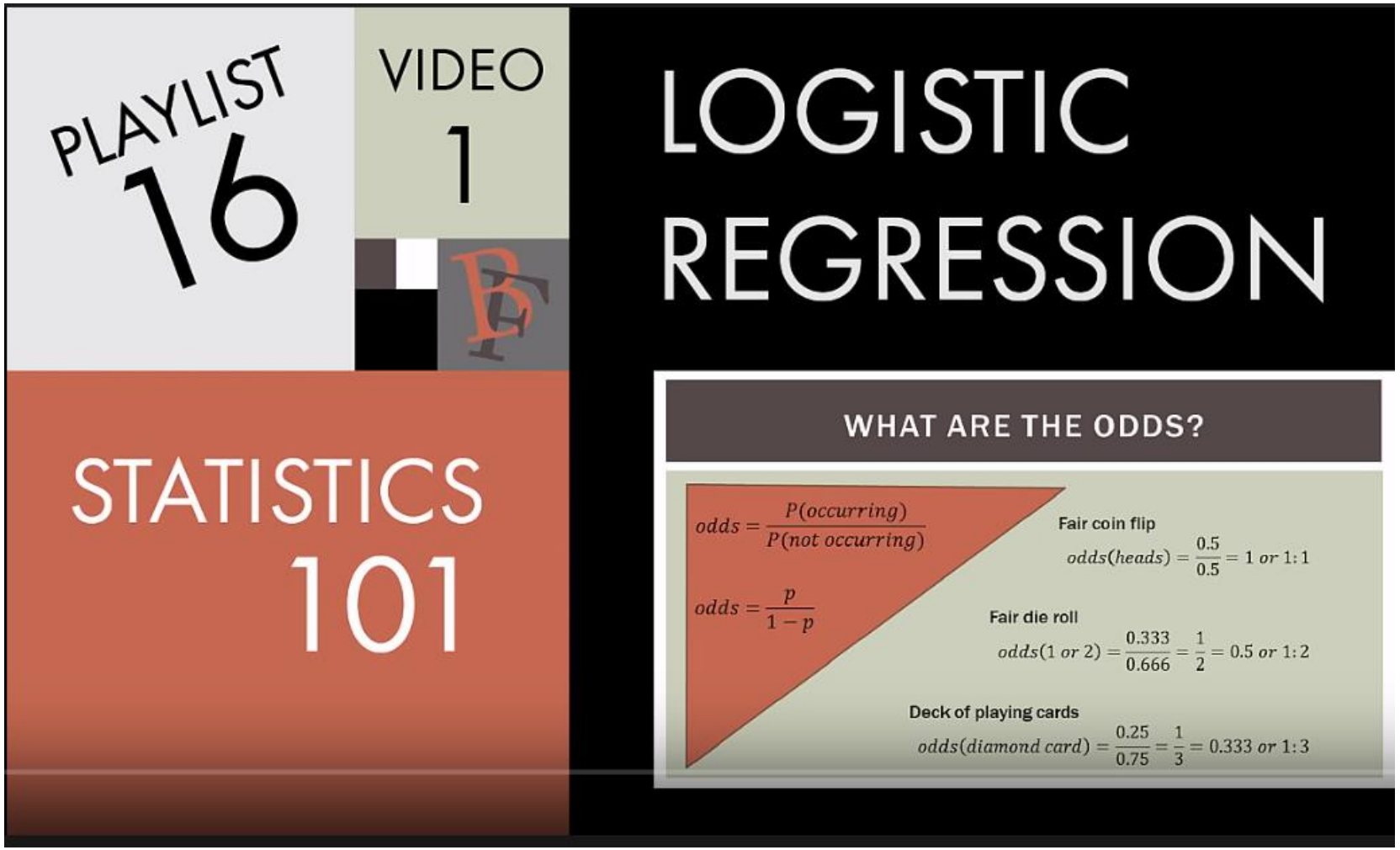
StatQuest – Regressão Logística



https://www.youtube.com/playlist?list=PLblh5JKOoLUKxzEP5HA2d-Li7IJkHfXSe

Escola Livre de IA

Brandon Foltz – Regressão Logística



https://youtu.be/zAULhNrnuL4

Escola Livre de IA

Escola Livre de IA

Escola Livre de IA