

MBA EM **ENGENHARIA
DE SOFTWARE**

**Técnicas de Machine
Learning**

Prof. Dr. Wilson Tarantin Junior

MBAUSP
ESALQ

A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização.

Lei nº 9610/98

Michel Ribeiro Corrêa 111.965.766-02

Machine learning

- Contextualização
 - Processamento dos dados por meio de algoritmos para a produção de informações
 - Aplica-se o machine learning com o intuito de **encontrar padrões nos dados**, seja criando modelos preditivos ou realizando análises exploratórias visando à busca de relações latentes
 - Envolve conhecimentos sobre estatística, softwares ou linguagem de programação e sobre o negócio
 - Frequentemente, o objetivo é a geração de informações úteis para fundamentar a tomada de decisão (*data-driven decision making*)

Banco de dados

- Definição e composição
 - O banco de dados armazena as informações de interesse para a análise em questão
 - Normalmente, o banco de dados contém uma amostra, ou seja, um subconjunto extraído da população
 - O banco de dados é composto por variáveis e por observações
 - **Observações:** unidades que têm suas características e atributos medidos
 - **Variáveis:** características/atributos observados, medidos ou categorizados

Banco de dados

- Bancos de dados sobre:

- Pessoas
- Países
- Empresas
- Tarefas
- Produtos
- Projetos
- ...

Michel Ribeiro Corrêa 111.965.766-02

Banco de dados

- Estrutura para uso
- Para aplicação em machine learning, comumente o banco de dados é estruturado com as variáveis em colunas e as observações em linhas em uma **estrutura tabular**

ID	Idade	Profissão	Renda Mensal	Estado (UF)	Escolaridade	...
Pessoa 1						
Pessoa 2						
Pessoa 3						
Pessoa 4						
Pessoa 5						
Pessoa 6						
Pessoa 7						
Pessoa n						

Tipos de variáveis

- As variáveis podem ser divididas em:
 - **Métricas:** são as variáveis quantitativas, isto é, apresentam características que podem ser mensuradas ou contadas
 - **Não métricas:** são as variáveis qualitativas, indicam as características que não podem ser medidas. Tais variáveis contêm categorias, por isto, muitas vezes, são chamadas de variáveis categóricas
 - **A identificação do tipo de variável é fundamental para a escolha da técnica que será utilizada na análise dos dados**

Tipos de variáveis: exemplos ilustrativos

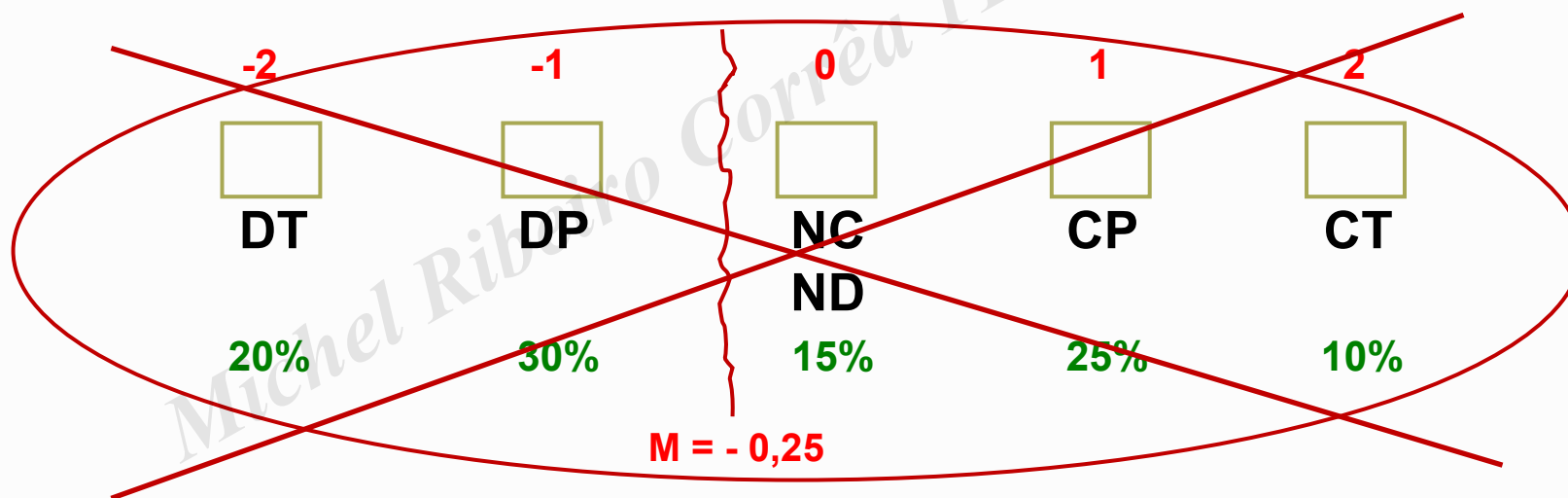
- Métricas
 - Idade – em anos
 - Renda mensal – em R\$
 - Número de habitantes no município – contagem
 - Quantidade de pessoas na equipe – contagem
- Categóricas (qualitativas)
 - Nacionalidade
 - Grau de escolaridade
 - Escalas likert
 - Linguagem de programação do projeto

Variáveis qualitativas

- Características principais
 - As variáveis qualitativas têm sua representação feita por tabelas de frequências ou gráficos formados a partir delas
 - **Tabela de frequências:** apresenta as contagens observadas por categoria da variável
 - Não é possível obter medidas descritivas como média ou desvio padrão para variáveis qualitativas

Variáveis qualitativas

- Um **erro** comum no tratamento de variáveis qualitativas é realizar **ponderação arbitrária**
- **Problema da ponderação arbitrária: gera resultados enviesados**



Ponderação Arbitrária !

Variáveis quantitativas

- Características principais
 - As variáveis quantitativas podem ser representadas por muitas ferramentas descritivas, como medidas de posição, dispersão e gráficos
 - A seguir, alguns exemplos comuns de estatísticas descritivas para variáveis métricas:
 - **Medidas de posição:** média, mediana, quartis
 - **Medidas de dispersão:** variância e desvio padrão

Modelos Lineares de Regressão Simples e Múltipla

Michel Ribeiro *Carreira* 111.965.766-02

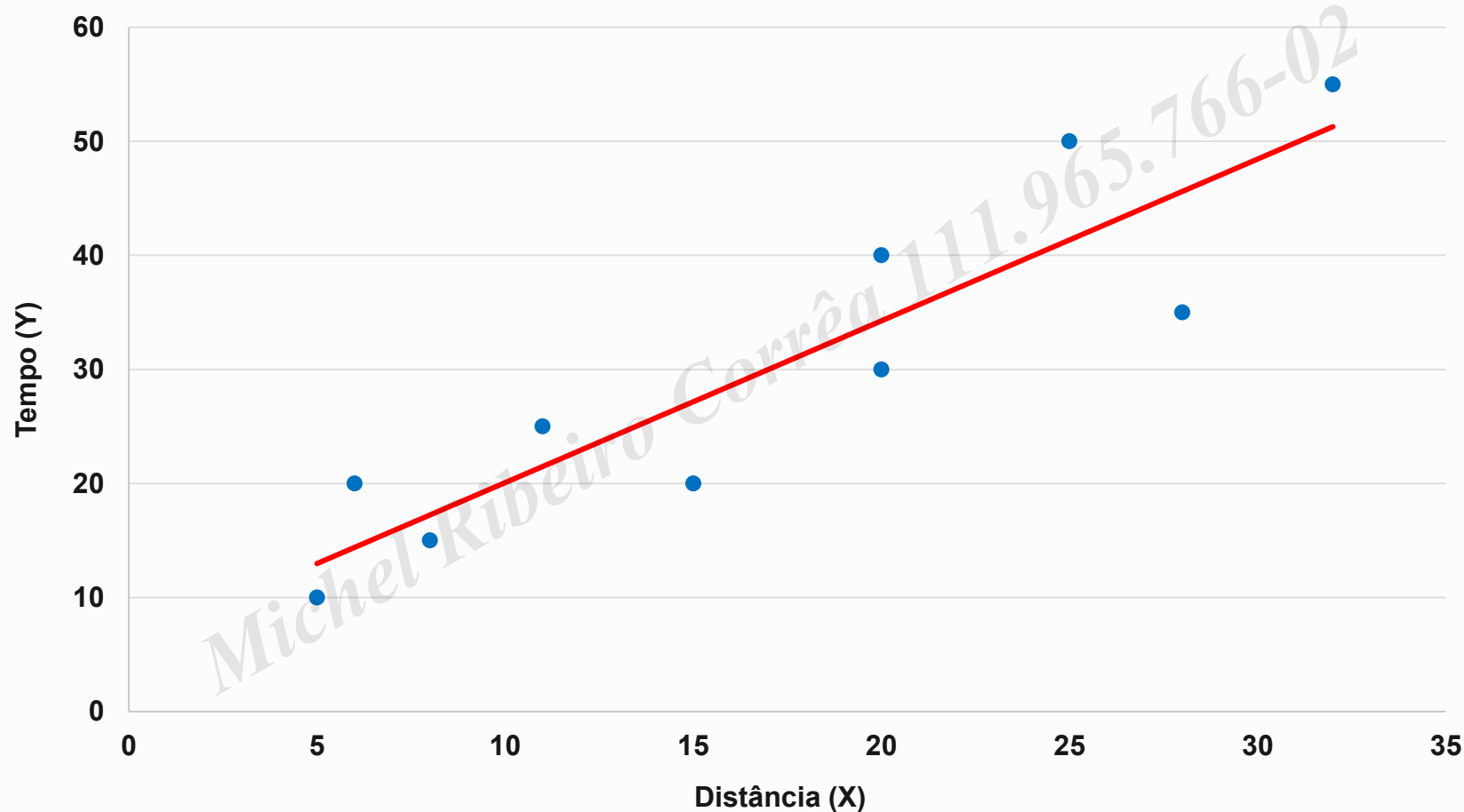
Modelos lineares de regressão

- Modelos supervisionados de machine learning
 - Conhecidos como modelos confirmatórios ou técnicas de dependência
 - O objetivo é estimar modelos, equações, com o intuito de elaborar previsões
 - Portanto, **há inferência dos resultados** para outras observações de fora da amostra
 - Define-se uma relação $Y = f(X_1, X_2, \dots, X_k)$
 - **Y**: chamada de variável dependente, é a variável a ser explicada no modelo (*target*)
 - **X**: chamadas de variáveis explicativas, são as preditoras (*features*)

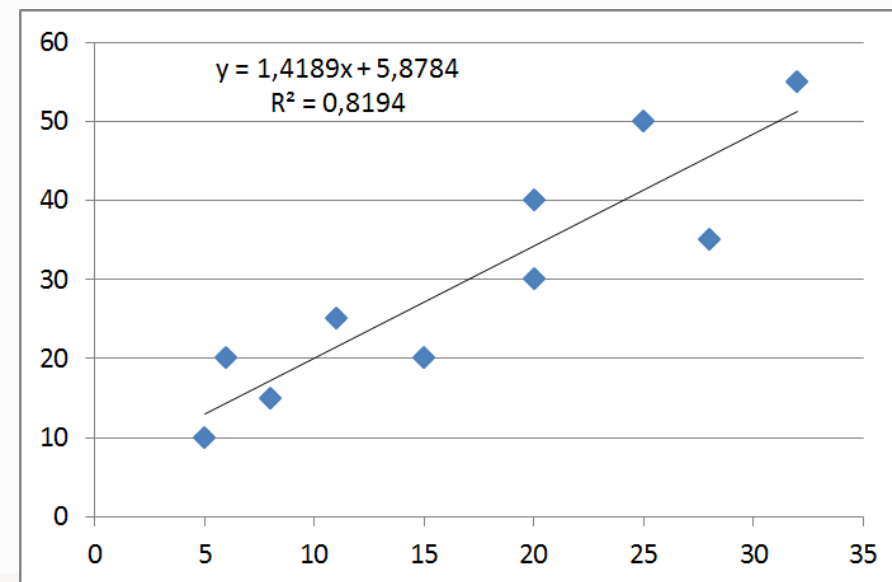
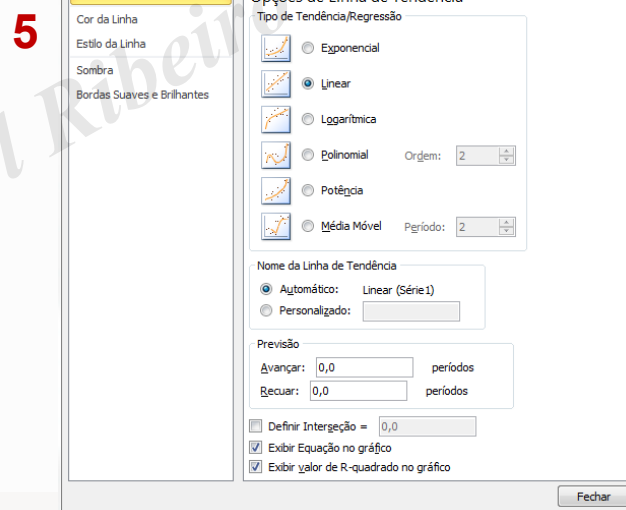
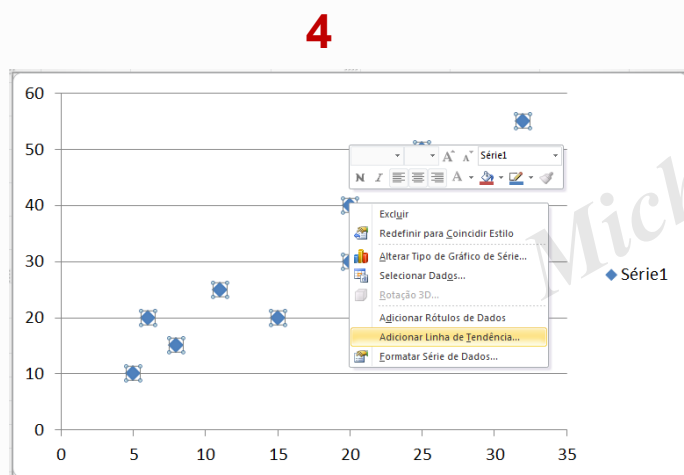
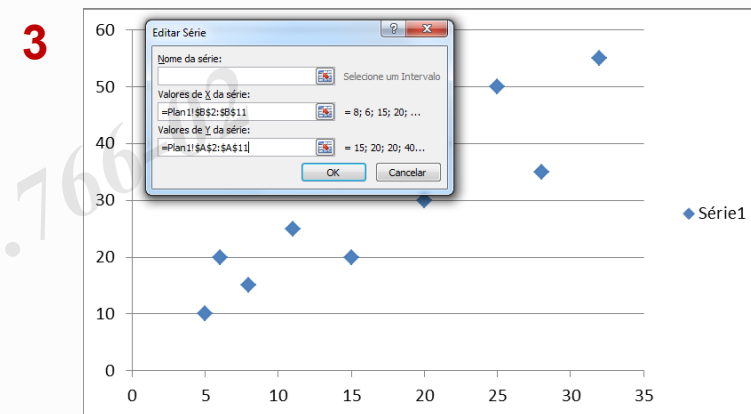
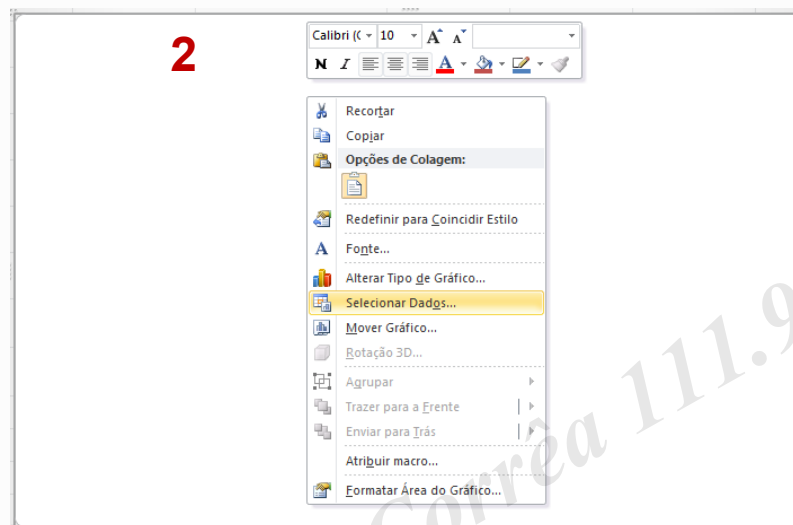
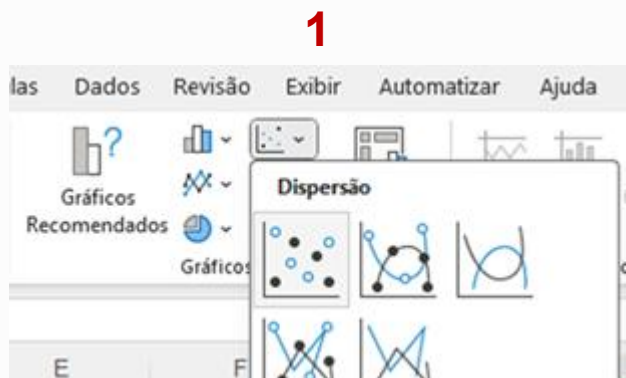
Quando aplicar o modelo

- A regressão linear é aplicada **quando a variável dependente é quantitativa**
 - O objetivo é explicar o comportamento de Y em função de um conjunto de X
 - Estabelece-se uma relação linear entre as variáveis
 - Regressão linear simples e múltipla:
 - A regressão linear simples contém apenas uma variável explicativa
 - A regressão linear múltipla contém mais de uma variável explicativa

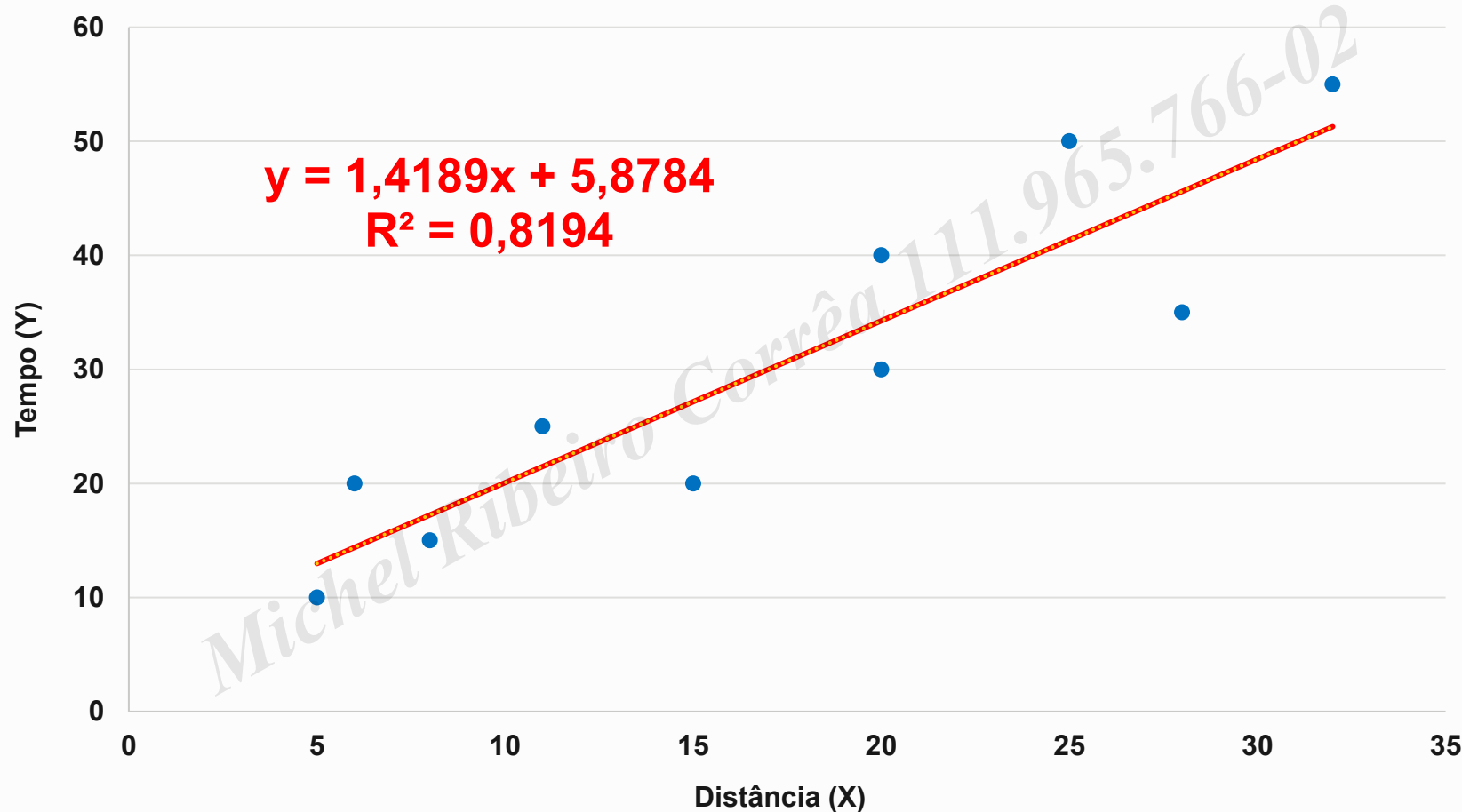
Visualizando graficamente um modelo linear de $Y = f(X)$



Qual é a equação que define o modelo?



Qual é a equação que define o modelo?



Modelo geral de regressão linear

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i$$

- **Y** é a variável dependente quantitativa (fenômeno em estudo)
- **a** representa a constante (intercepto)
- **b_j** representam os coeficientes para cada variável explicativa
- **X_j** representam as variáveis explicativas do modelo
- **u_i** representa o termo de erro do modelo
- **k** é o número de variáveis explicativas e **i** refere-se às observações em análise
- **As variáveis explicativas (X) podem ser métricas ou categóricas**

Mínimos quadrados ordinários (MQO)

- Serão estimados os parâmetros α e β do modelo, sendo \hat{Y}_i o valor previsto por observação

$$\hat{Y}_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \cdots + \beta_k \cdot X_{ki}$$

- Sendo assim, é possível definir o resíduo do modelo para dada observação i como:

$$u_i = Y_i - \hat{Y}_i$$

- Condições para a estimação dos parâmetros do modelo (MQO)
 - 1. A somatória dos resíduos deve ser igual a zero
 - 2. A somatória dos resíduos ao quadrado é a mínima possível

Otimização de
P.O. (Exemplo
Solver Excel)

Elementos de um modelo

- Interpretaremos:
 - Coeficientes estimados
 - Significância geral do modelo (teste F – ANOVA)
 - Significância dos parâmetros (testes t)
 - Intervalos de confiança
 - Poder explicativo do modelo (R^2)

Michel Ribeiro Corrêa 111.965.766-02

Parâmetros do modelo

- Interpretação dos parâmetros estimados α e β
 - α é o coeficiente linear (o intercepto), ou seja, o valor de Y caso todas as X = 0
 - Muitas vezes, o α pode ser interpretado como a projeção da reta no eixo Y, uma vez que não encontram-se observações da amostra com todas as variáveis X = 0
 - β são os coeficientes angulares, ou seja, a inclinação da reta
 - Na regressão múltipla, os β são interpretados na condição *ceteris paribus*, ou seja, o efeito de certa variável X sobre Y mantidas todas as demais variáveis X constantes
 - A interpretação dos parâmetros do modelo deve ocorrer **sem a extrapolação** dos dados, isto é, são válidos dentro do limite de variação das variáveis X na amostra

Teste F (ANOVA)

- Significância geral do modelo: testa se pelo menos um dos β estimados é estatisticamente diferente de zero

$$F = \frac{\frac{SQR}{(k-1)}}{\frac{SQU}{(n-k)}}$$

SQR: Soma dos Quadrados da Regressão
SQU: Soma dos Quadrados dos Resíduos
k: nº de parâmetros do modelo (incluindo α)
n: tamanho da amostra

- $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
- H_1 : existe pelo menos um $\beta_j \neq 0$
- Normalmente, adota-se o nível de significância de 5% para o teste
- **Se o p-valor do teste $F < 0.05$, rejeita-se H_0**

Teste F (ANOVA)

$$\text{SQT} = \text{SQR} + \text{SQU}$$

- **Soma dos Quadrados Totais (SQT):** variação de Y em torno de sua média
- **Soma dos Quadrados da Regressão (SQR):** variação de Y explicada pelas variáveis X
- **Soma dos Quadrados dos Resíduos (SQU):** variação de Y não explicada pelo modelo

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{SQT}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{SQR}} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{\text{SQU}}$$

Teste t

- Avalia a significância individual dos parâmetros estimados

$$t_{\alpha} = \frac{\alpha}{s.e.(\alpha)} \quad t_{\beta_j} = \frac{\beta_j}{s.e.(\beta_j)}$$

- $H_0: \alpha = 0$
- $H_1: \alpha \neq 0$
- $H_0: \beta_j = 0$
- $H_1: \beta_j \neq 0$
- Comumente, adota-se o nível de significância de 5% para o teste
- Se o p-valor do teste $t < 0.05$, rejeita-se H_0**
- Mesmo que não tenha significância, o α não deve ser removido do modelo!**

Intervalos de confiança

- Para dado nível de confiança, é o intervalo de valores que contém o verdadeiro parâmetro populacional

$$\alpha \pm t \times s.e.(\alpha)$$

$$\beta_j \pm t \times s.e.(\beta_j)$$

- t é o valor crítico bicaudal da distribuição t de Student para o nível de confiança escolhido na análise, com $n - k$ graus de liberdade
- Normalmente, observa-se o nível de confiança de 95% (nível de significância de 5%)

Coeficiente de explicação (R^2)

- O R^2 apresenta o poder explicativo do modelo, ou seja, o percentual da variabilidade de Y que é explicado pela variação conjunta das variáveis X

$$R^2 = \frac{SQR}{SQR + SQU}$$

- **O R^2 varia de 0 a 1: valores mais próximos de 1 indicam maior capacidade preditiva**
 - O R^2 não deve ser analisado no sentido de validar ou não o modelo, pois, em muitos campos do conhecimento, é comum não obter valores muito elevados
 - R^2 ajustado para comparação entre modelos: $R^2_{ajust} = 1 - \frac{n-1}{n-k} (1 - R^2)$
 - Ajusta-se a quantidade k de parâmetros (incluindo o α) e o tamanho da amostra n

Variáveis explicativas categóricas

- Quando há variáveis X categóricas, é necessário transformá-las em **dummies**
- Dummy**: variável binária (1 ou 0) indicando a presença (1) ou ausência (0) do atributo

ID	Variável A		Variável B		
	Categ. 1	Categ. 2	Categ. 1	Categ. 2	Categ. 3
1	1	0	0	1	0
2	0	1	0	0	1
3	0	1	1	0	0
4	1	0	0	0	1
5	0	1	0	1	0
6	1	0	1	0	0

Na regressão, utiliza-se o procedimento de **$n - 1$ dummies**, isto é, uma das categorias de cada variável categórica deve ficar como a referência de sua variável no intercepto

Referência / Sugestão de Leitura

- Fávero, Luiz Paulo; Belfiore, Patrícia. (2024). Manual de análise de dados: estatística e machine learning com Excel®, SPSS®, Stata®, R® e Python®. 2 ed. Rio de Janeiro: LTC.
- Wooldridge, Jeffrey M. (2017). Introdução à Econometria: uma abordagem moderna. 3 ed. Editora Cengage.

Michel Ribeiro Corrêa 111.965.766-32



Stats Studio

Acesse: stats-studio.com

Michel Ribeiro Corrêa 11.965.766-02

Clique em “**Registrar**”

STATS studio

Planos

Acessar plataforma

Registrar

Ciência de dados online e sem código

Explore, analise e modele seus dados para
descobrir insights de forma simples e intuitiva,
sem digitar uma linha de código!

Comece gratuitamente

Aprenda como



Stat's Studio

Criar Conta

Usuário

CRIE UM USUÁRIO

E-mail

AQUI É SEU E-MAIL

Digite um e-mail válido. Essa informação será utilizada para recuperação de senha.

Senha

CRIE UMA SENHA

Mínimo de 6 dígitos

Confirmar senha

REDIGITE A SENHA CRIADA

Cadastrar

Já tem uma conta? [Faça seu login aqui!](#)

Stat's Studio

Login

Usuário

USUÁRIO CRIADO

Senha

SENHA CRIADA

☒ Lembrar-me

[Esqueceu a senha?](#)

Acessar

Ainda não tem uma conta? [Registre aqui!](#)

Criando um projeto: **File** → **New project**



Criando um projeto: **nome, tipo de arquivo e importar**

New project

Project name *

modelo_reg

Select database *

Importar arquivo Excel

Import dataset *

Escolher arquivo tempodist.xls

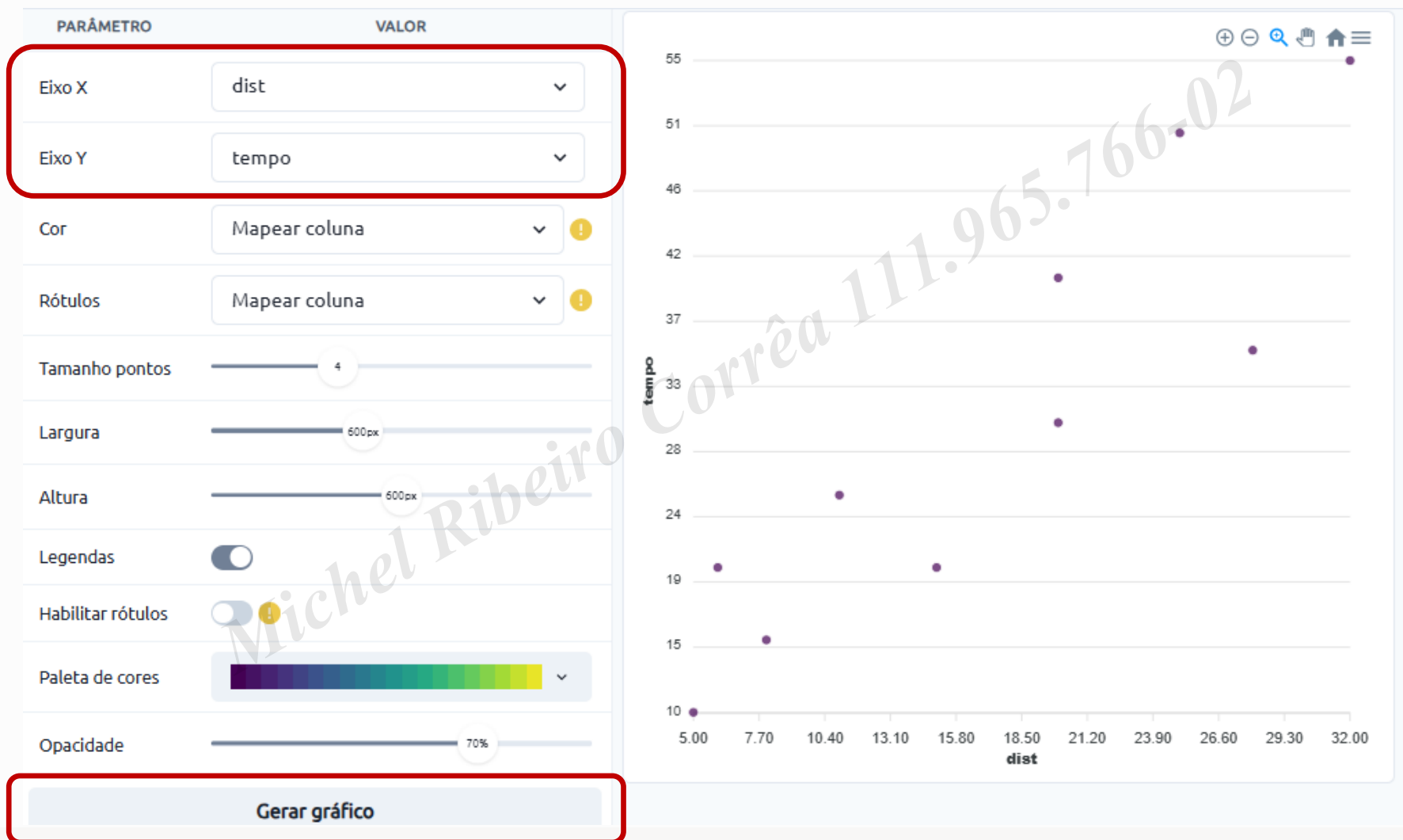
Cancel

Confirm

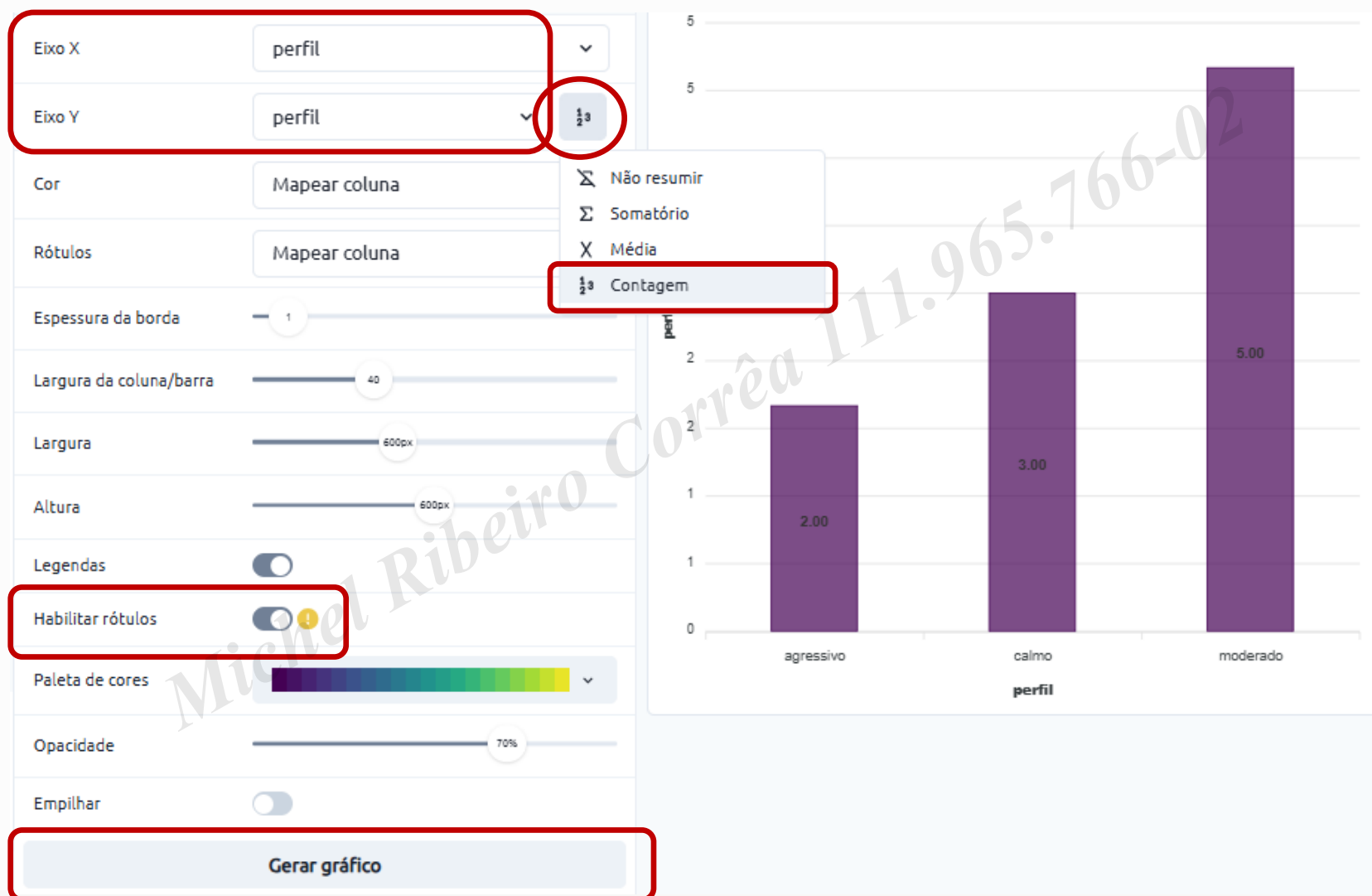
Estatísticas descritivas: **Describe dataset e Correlation Matrix**

File	Dataset	Upgrade	US
Dataset	Project: modelo_reg	Mostrando todas as 5 colunas	
Data Wrangling	Rows count: 10		
Table	ESTUDANTE	TEMPO	DIST
fx Add calculated column	TIPO: Texto	TIPO: Inteiro	TIPO: Inteiro
Group by and summarize	MISSING: 0 (0.00)%	MISSING: 0 (0.00)%	MISSING: 0 (0.00)%
Auxiliary datasets	Patricia	15	8
Analysis	Gabriela	20	6
Describe dataset	Luiz Felipe	20	15
Correlation Matrix	Ovidio	40	20
Modelagem	Leonor	50	25
Supervised	Leticia	25	11
Unsupervised	Gustavo	10	5
Time series	Dalila	55	32

Data visualization: Gráficos → Pontos



Realize data visualization: Gráficos → Coluna



Gere o modelo de regressão: **Supervised**

File	Dataset	Upgrade	US
Dataset	Project: modelo_reg		
Data Wrangling	Rows count: 10		Mostrando todas as 5 colunas
Table	ESTUDANTE	TEMPO	DIST
fx Add calculated column	TIPO: Texto	TIPO: Inteiro	TIPO: Inteiro
Group by and summarize	MISSING: 0 (0.00)%	MISSING: 0 (0.00)%	MISSING: 0 (0.00)%
Auxiliary datasets	Patricia	15	8
Analysis	Gabriela	20	6
Describe dataset	Luiz Felipe	20	15
Correlation Matrix	Ovidio	40	20
Modelagem	Leonor	50	25
Supervised	Leticia	25	11
Unsupervised	Gustavo	10	5
Time series	Dalila	55	32

Gere o modelo de regressão: tipo de modelo e variáveis

Tipo de modelo	Variável dependente	Variáveis explicativas	Resetar seleção	Selecionar todas
Regressão Linear	tempo	Selecione as varáveis: estudante tempo dist sem perfil		
Fórmula				
$\text{tempo}_i = \alpha + \beta_1 \cdot \text{dist}_i + \mu_i$				
Parâmetros				
<input type="checkbox"/> Stepwise	<input type="checkbox"/> Visualizar betas padronizados			
<div>Simular</div>				

Novo experimento

Nome

reg_inicial

Cancelar

Confirmar

Interpretando os resultados do modelo

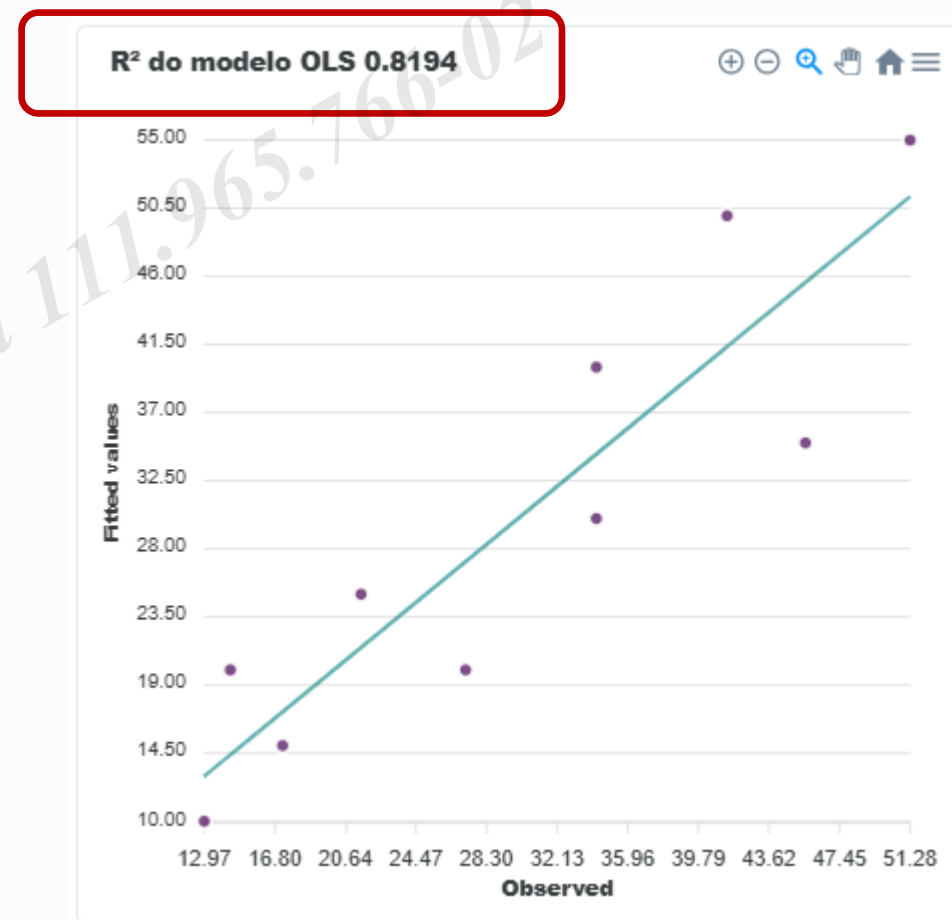
PARÂMETRO	VALOR	ERRO PADRÃO	t	P-VALUE	[2.5%	97.5%]
const	5.87838	4.53233	1.29699	0.23079	-4.57319	16.32994
dist	1.41892	0.23550	6.02520	0.00031 ***	0.87586	1.96198

alfa e beta

significância

$$\widehat{tempo}_i = 5,87 + 1,41 \cdot distância_i$$

poder explicativo



Transforme variáveis categóricas para *dummies*

1

PERFIL

GERAL

Renomear Coluna

TRATAMENTOS

Encodar rótulos

Dummizar variável (n)

2

Dummizar variável (n-1)

Excluir coluna

calmo

moderado

moderado

Dummizar variável: perfil

Categoria de referência

Selecione a categoria

Selecione a categoria

agressivo

calmo

moderado

3

Gerando novo modelo com *dummies*

Tipo de modelo	Variável dependente	Variáveis explicativas	Resetar seleção	Selecionar todas
Regressão Linear ▾	tempo ▾	Selecione as variáveis: estudante tempo dist sem perfil_orig perfil_agressivo perfil_moderado		
Fórmula				
$\text{tempo}_i = \alpha + \beta_1.\text{dist}_i + \beta_2.\text{sem}_i + \beta_3.\text{perfil_agressivo}_i + \beta_4.\text{perfil_moderado}_i + \mu_i$				
Parâmetros				
<input type="checkbox"/> Stepwise	<input type="checkbox"/> Visualizar betas padronizados			
Simular				

Novo experimento

Nome

reg_dummies

Cancelar

Confirmar

Interpretando os resultados do modelo

PARÂMETRO	VALOR	ERRO PADRÃO	t	P-VALUE	[2.5%	97.5%]
const	8.18358	1.08082	7.57162	0.00064 ***	5.40524	10.96193
dist	0.70775	0.07427	9.52973	0.00022 ***	0.51684	0.89867
sem	7.86676	0.74598	10.54554	0.00013 ***	5.94916	9.78436
perfil_agressivo	9.09179	1.28670	7.06596	0.00088 ***	5.78421	12.39937
perfil_moderado	0.19893	1.00692	0.19756	0.85117	-2.38944	2.78730

alfa e betas

significâncias

! Seu modelo pode não ser estatisticamente significativo!
Ele possui pelo menos um beta com p-value maior que 0.05

Gerando novo modelo com *Stepwise*

Tipo de modelo	Variável dependente	Variáveis explicativas	Resetar seleção	Selecionar todas
Regressão Linear	tempo	Selecione as variáveis: estudante tempo dist sem perfil_orig perfil_agressivo perfil_moderado		
Fórmula				
$\text{tempo}_i = \alpha + \beta_1.\text{dist}_i + \beta_2.\text{sem}_i + \beta_3.\text{perfil_agressivo}_i + \beta_4.\text{perfil_moderado}_i + \mu_i$				
Parâmetros				
<input checked="" type="checkbox"/> Stepwise <input type="checkbox"/> Visualizar betas padronizados				
Simular				

Novo experimento

Nome

reg_final

Cancelar

Confirmar

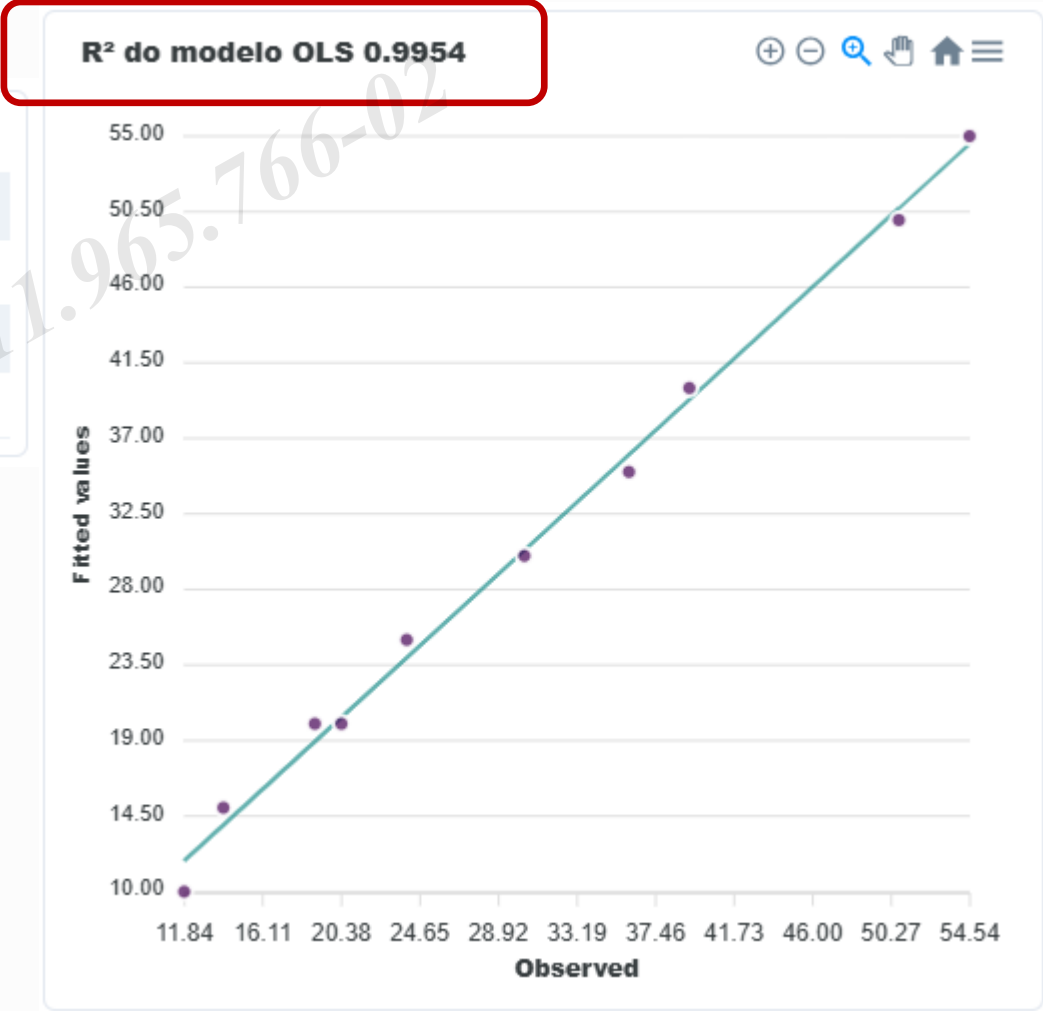
Interpretando os resultados do modelo

PARÂMETRO	VALOR	ERRO PADRÃO	t	P-VALUE	[2.5%	97.5%]
const	8.29193	0.85351	9.71512	0.00007 ***	6.20347	10.38039
dist	0.71045	0.06690	10.61953	0.00004 ***	0.54675	0.87415
sem	7.83684	0.66940	11.70721	0.00002 ***	6.19887	9.47481
perfil_agressivo	8.96761	1.02889	8.71580	0.00013 ***	6.45000	11.48521

alfa e betas

significâncias

poder explicativo



Modelo preditivo final: realizar previsões

$$\widehat{tempo}_i = 8,29 + 0,71 \cdot dist_i + 7,83 \cdot sem_i + 8,96 \cdot perfil_agressivo_i$$



Prever resultado

dist	sem	perfil_agressivo
30	2	1

Prever

Resultado: 54.246820349761535

A dashed green arrow points from the left towards the 'dist' input field. A red box highlights the 'Prever' button. A green box highlights the 'Resultado' output field.

Obrigado!

Prof. Wilson Tarantin Junior | [linkedin.com/in/wilson-tarantin-junior-359476190](https://www.linkedin.com/in/wilson-tarantin-junior-359476190)

MBAUSP
ESALQ