24.07.365

CONQUER

e-book

Pós-Graduação em Business Intelligence & Analytics.

Como alcançar
resultados de
forma inteligente.

jornada do curso

MÓDULO 01_O PRINCÍPIO DE TUDO

MÓDULO 02_DADOS, INFORMAÇÕES E INTELIGÊNCIA

MÓDULO 03_OS DADOS CONTAM HISTÓRIAS

AULA 07_ CONECTAR DADOS A RESULTADOS: OLHAR PARA TRÁS

→ **AULA 08_CONECTAR DADOS A RESULTADOS: OLHAR PARA FRENTE**

AULA 09_DADOS COMO HIPÓTESES NA NOVA ECONOMIA: ENTRE PREVISÕES E RESULTADOS DE SUCESSO

AULA 10_CASE #01

AULA 11_CHECKPOINT

MÓDULO 04_APRESENTAR OS DADOS E AS DECISÕES

MÓDULO 03:
OS DADOS CONTAM HISTÓRIAS

CONECTAR DADOS A RESULTADOS: OLHAR PARA FRENTE

ou vai ou voa

PÓS-GRADUAÇÃO EM
BUSINESS INTELLIGENCE & ANALYTICS

sumário

07 conceitos estatísticos

11 correlação

18 regressão linear

34 #desafioconquer08

36 #queromais

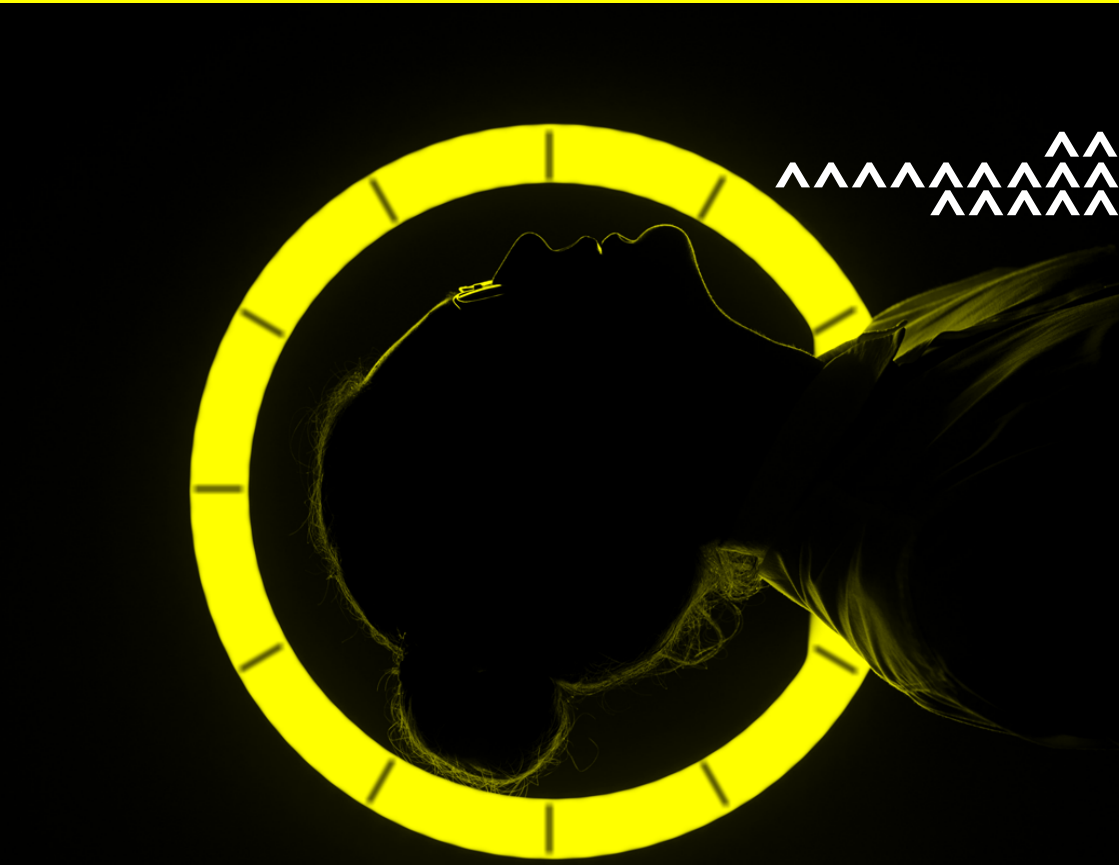
Clique, **selecione** ou **escreva**! Este material possui recursos interativos para enriquecer sua experiência.

MISSÃO

missão da aula

08

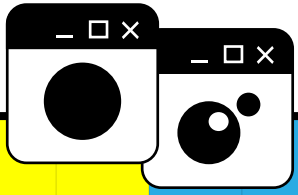
Entender **como** o **passado** nos ajuda a tomar melhores decisões, por meio das **análises preditivas**.



a análise do **passado** garante uma melhor avaliação e **previsão dos resultados** do futuro.

conceitos estatísticos

conceitos estatísticos



POPULAÇÃO

Conjunto de indivíduos que compartilham, pelo menos, uma característica em comum.

exemplos:
Etnia, matrícula na universidade, tipos de uma fruta ou verdura, carros em Curitiba, carros de uma marca etc.

AMOSTRA

Subconjunto de indivíduos extraídos da população.

exemplos:

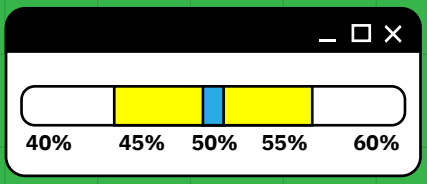
- Analisar uma parte da população de certa etnia para descobrir a faixa etária predominante.
- Investigar parte da população dos alunos matriculados na universidade para saber de onde esses estudantes vieram.
- Examinar a amostra dos carros de uma marca para saber qual é a quilometragem média que andam por ano.

conceitos estatísticos

MARGEM DE ERRO

Trata-se da diferença entre o **resultado obtido** com a **amostra** e o **resultado real da população**.

exemplo: Se a margem de erro de uma pesquisa é de 5%, isso significa que se 50% dos entrevistados (amostra) gostam de um produto, você deve considerar que esse número, na população, pode oscilar entre 45% e 55%.



NÍVEL DE CONFIANÇA

É o nível de certeza de que os dados medidos refletem a população, ou seja, ele garante que o valor exato está dentro da margem de erro.

exemplo: Se o nível de confiança de uma pesquisa é de 98%, isso significa que, se ela for aplicada 100 vezes, ela daria resultados dentro da margem de erro em 98 casos.

Geralmente, utiliza-se o nível de confiança de 95%, mas podem ser usados outros níveis de confiança como 90%, 99% ou outro, dependendo somente do quão importante é para o pesquisador que o valor esteja o mais próximo do valor real da população.

EXEMPLO:
Com 95% de confiança, podemos afirmar que 50% dos clientes preferem o produto A, com uma margem de erro de 3% para cima ou para baixo.

OU SEJA:
Temos 95% de certeza que entre 47 e 53% dos clientes preferem o produto A.

conceito estatístico

NÍVEL DE SIGNIFICÂNCIA:

É conhecido como **ALFA**. Ele indica a probabilidade de que o valor encontrado não está dentro da margem de erro esperada.

ALFA = 100% - nível de confiança
Como é utilizado o nível de confiança de 95%, o valor do alfa, em geral, será de 5%.

cor-relação

correlação

Significa uma semelhança ou relação entre **duas coisas, pessoas ou ideias.**

Na estatística, ela nos ajuda a determinar qual é a intensidade da relação que existe entre 2 variáveis:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

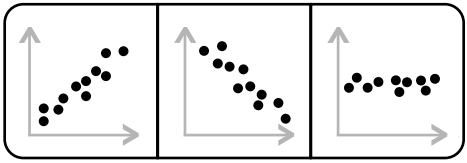
EXCEL

=CORREL(X;Y)

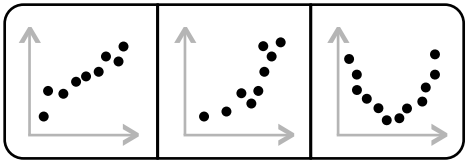
PHYTON

X.corr(Y)

tipos de correlação



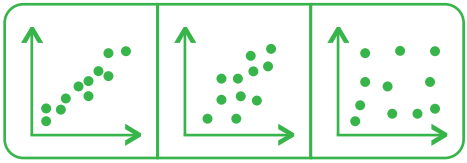
positiva negativa nula



linear exponencial em forma de U

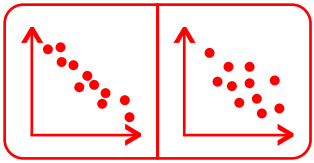
força da correlação

POSITIVA



forte(~0,9) fraca(~0,3) nenhuma(~0)

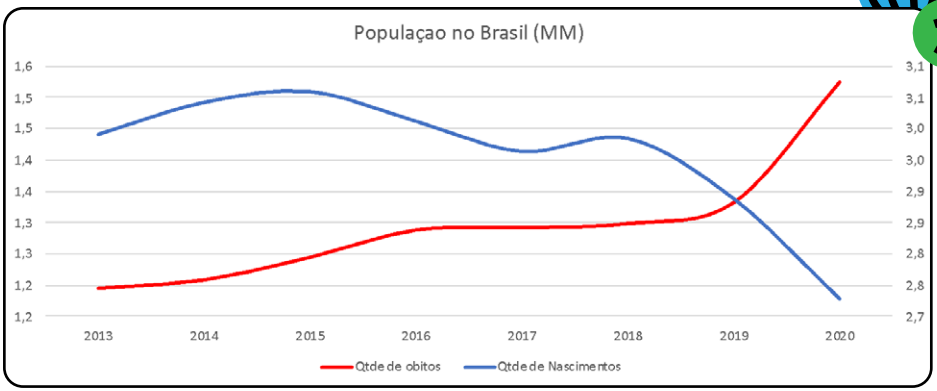
NEGATIVA



forte(~ -0,9) fraca(~ -0,3)

exemplos de correlação:

De acordo com o IBGE, temos:



Correlação: -0,93

Ou seja, quanto menos nascimentos, mais óbitos.
Portanto, se nascer muitas pessoas, não haverá mais óbitos.

“O galo sempre canta antes de nascer o sol. Logo, o sol nasce porque o galo canta.”

Na verdade, o galo canta para avisar o galinheiro que continua vivo e no comando; para demarcar território e indicar que quem manda naquele espaço é ele.

Ou seja, o que acontece aqui é a suposição de que uma relação real – ou percebida entre duas coisas – significa que uma é a causa da outra.

correlação não implica em causalidade!

A correlação nos diz a força e a direção do relacionamento entre variáveis, mas nada esclarece sobre os motivos desse relacionamento.

mão na massa

RH - parte 1

Você trabalha no RH e tem um candidato com 5 anos de experiência.
Qual é o melhor salário que você deve oferecer a ele?

- Existe correlação entre: **anos de experiência x salário**?
- Faça um **diagrama de dispersão** - parece existir uma correlação linear entre as variáveis?
- Calcule o **coeficiente de correlação**.

Importante: baixe o arquivo **salário.xlsx** disponível para download **clicando aqui**.

resposta abaixo (passe o mouse)

mão na massa

Imobiliária - parte 1

A imobiliária House solicitou um modelo para estimar o preço dos imóveis de acordo com a área do terreno. Existe correlação entre: **preço x área**?

> Faça um **diagrama de dispersão** - parece existir uma correlação linear entre as variáveis?

> Calcule o **coeficiente de correlação**.

Importante: baixe o arquivo **imobiliária.xlsx** disponível para download **clicando aqui**.

resposta abaixo (passe o mouse)

regressão linear

regressão linear simples

Utilizamos a regressão linear simples para descrever a relação linear entre duas variáveis, ou seja, quando queremos prever o valor de uma variável utilizando apenas uma outra variável.

Assim, temos:
> uma variável dependente Y, ou **resposta**.
> uma independente X, também conhecida como **explicativa**.

reta de regressão linear

A relação entre as 2 variáveis pode ser descrita por meio de uma função linear:

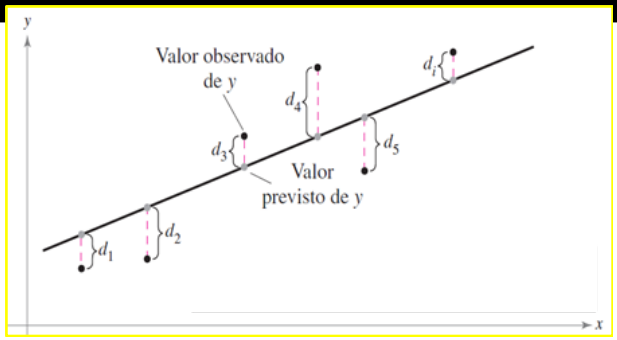
$$\hat{y}_i = b_o + b_1 x_i$$

$\hat{y}_i$: variável dependente (resposta)

b_o: intercepto (coef. linear)

b_1: inclinação da reta (coef. angular)

x_i: variável independente (explicativa)



mão na massa

RH - parte 2

Você trabalha no RH e tem um candidato com 5 anos de experiência.
Qual é o melhor salário que você deve oferecer a ele?

- Encontre a **reta de regressão** para os dados do arquivo
- Qual o salário para o candidato com 5 anos de experiência ?
- Continue utilizando o arquivo **salário.xlsx.**

resposta abaixo (passe o mouse)

regressão linear simples

método 01:
Gráfico de dispersão + linha de tendência.

passo a passo:

- Criar um gráfico de dispersão;
- Clicar com o botão direito nos dados da série;
- Selecionar: adicionar linha de tendência;
- Opção da linha de tendência: LINEAR;
- Selecionar: exibir equação no gráfico.

método 02:
Fórmulas

passo a passo:

- Utilizar as fórmulas para calcular os coeficientes:
  b_0=INTERCEPÇÃO(Y,X)
  b_1=INCLINAÇÃO(Y,X)

método 03:
Ferramenta de análise de dados.

passo a passo:

- Habilitar o suplemento: ferramentas de análise;
- Selecionar a aba: DADOS;
- Selecionar a opção: análise de dados;
- Ferramenta de análise: regressão;
- Preencher com as informações da tabela.

regressão linear simples

Utiliza dados do passado para prever comportamentos e detectar padrões no conjunto de dados analisados.

Seu objetivo é se aprofundar no que aconteceu no **passado** para obter uma melhor avaliação e previsão dos **resultados do futuro.**

vantagens:
01. Detecta fraudes;
02. Otimiza campanhas de marketing;
03. Melhora operações;
04. Reduz riscos;
05. Torna melhor a gestão de clientes.

mão na massa

Imobiliária - parte 2

A imobiliária House solicitou um modelo para estimar o preço dos imóveis de acordo com a área do terreno.

- Encontre a **reta de regressão** para os dados do arquivo.
- Qual o **preço** para um imóvel com **área** igual à 2150? E 3280?
- Continue utilizando o arquivo **imobiliária.xlsx.**

resposta abaixo (passe o mouse)

a equação encontrada se ajusta bem aos nossos dados?

É importante checar as medidas de regressão e os resíduos do modelo para responder a essa pergunta.

entendendo parâmetros

Quão bem a equação de **regressão linear** encontrada **se encaixa em seus dados de origem?**

Estatística de regressão	
R múltiplo	0,97824162
Quadrado de R	0,95695666
Quadrado de R ajustado	0,9554194
Erro-padrão	5788,31505
Observações	30

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significância*
Regressão	1	2,0857E+10	2,0857E+10	622,507203	1,14307E-20
Residual	28	938128552	33504591,1		
Total	29	2,1795E+10			

	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*	*Inferior 95,0%*	*Superior 95,0%*
Interceptar	25792,2002	2273,05343	11,3469397	5,512E-12	21136,06131	30448,33908	21136,06131	30448,33908
Anos Experiência	9449,96232	378,754574	24,9500942	1,1431E-20	8674,118747	10225,8059	8674,118747	10225,8059

Estatística de regressão	
R múltiplo	0,97824162
Quadrado de R	0,95695666
Quadrado de R ajustado	0,9554194
Erro-padrão	5788,31505
Observações	30

> É o coeficiente de correlação, que já analisamos anteriormente.

Estatística de regressão	
R múltiplo	0,97824162
Quadrado de R	0,95695666
Quadrado de R ajustado	0,9554194
Erro-padrão	5788,31505
Observações	30

> É o coeficiente de determinação. Ele quer dizer o quanto da variação em Y é explicada por X, em outras palavras, ele nos mostra se o modelo possui um bom ajuste.
Ele varia de 0% a 100%: neste caso, 95% da variação no **preço** do aluguel é explicada pela variação da **área** (os outros 5% que faltam é explicado por outras variáveis que não foram consideradas no modelo + o erro amostral).

Estatística de regressão	
R múltiplo	0,97824162
Quadrado de R	0,95695666
Quadrado de R ajustado	0,9554194
Erro-padrão	5788,31505
Observações	30

> É o coeficiente de determinação ajustado para o número de variáveis independentes. Utilizamos para comparar modelos com diferentes quantidades de variáveis independentes.

Estatística de regressão	
R múltiplo	0,97824162
Quadrado de R	0,95695666
Quadrado de R ajustado	0,9554194
Erro-padrão	5788,31505
Observações	30

> Indica a precisão da análise em valores absolutos. Mostra, em média, o quanto a estimativa pode estar errada para mais ou para menos.

Estatística de regressão	
R múltiplo	0,97824162
Quadrado de R	0,95695666
Quadrado de R ajustado	0,9554194
Erro-padrão	5788,31505
Observações	30

> Quantidade de observações utilizadas para construção do modelo, isto é, o número de linhas na tabela origem.

ANOVA					
	gl	*SQ*	*MQ*	*F*	*F de significância*
Regressão	1	2,0857E+10	2,0857E+10	622,507203	1,14307E-20
Residual	28	938128552	33504591,1		
Total	29	2,1795E+10			

F de significância nos diz se o modelo é estatisticamente significante, por meio do teste F. Se esse valor for menor que o nível de significância, podemos concluir que o modelo estimado é estatisticamente significativo. Caso seja maior que o nível de significância, o ideal seria rever o seu modelo e a escolha das variáveis independentes que foram utilizadas.

	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*	*Inferior 95,0%*	*Superior 95,0%*
Interceptar	25792,2002	2273,05343	11,3469397	5,512E-12	21136,06131	30448,33908	21136,06131	30448,33908
Anos Experiência	9449,96232	378,754574	24,9500942	1,1431E-20	8674,118747	10225,8059	8674,118747	10225,8059

Valor P nos diz se cada uma das variáveis é estatisticamente significante, por meio do teste F.

Se esse valor for menor que nível de significância, podemos concluir que a variável é estatisticamente significativa. Caso seja maior que o nível de significância, o ideal seria retirar essa variável e recalcular o modelo

mão na massa

Imobiliária - parte 3

A imobiliária House solicitou um modelo para estimar o preço dos imóveis de acordo com a área do terreno.

> A reta de regressão está **bem ajustada**?

> Continue utilizando o arquivo **imobiliária.xlsx.**

regressão linear múltipla

É utilizada quando se quer investigar a relação entre uma variável dependente Y e duas ou mais variáveis independentes X's.
O modelo é representado por:

$$\hat{y} = b_0 + b_1x_1 + \ldots + b_nx_n$$

mão na massa

Carros - parte 1

Você foi contratado por uma empresa automobilística e eles vão lançar um novo modelo. Qual deve ser o **preço** deste carro?

- Calcule a **matriz de correlações.**
- Estime o **modelo de regressão**, encontrando a **equação de regressão** que se ajuste aos dados.
- Avalie as **medidas de ajuste** do modelo.
- Faça **novas estimativas.**

Importante: baixe o arquivo **carros.xlsx** disponível para download **clicando aqui**.

podemos melhorar o nosso modelo?

mão na massa

Carros - parte 2

Como vimos, o modelo utilizando todas as variáveis não é o melhor modelo, pois temos variáveis com **p-valor alto.**

> Agora, precisamos então recalcular o modelo de Regressão apenas com as **melhores variáveis**.

> **Quais variáveis** vocês selecionariam?

> **Recalcule o modelo** com estas variáveis.

> Continue utilizando o arquivo **carros.xlsx.**

pulo do gato

menos é **MAIS**!

Modelos mais simples (ou com menos variáveis) devem ser escolhidos desde que a qualidade do ajuste seja similar, ou seja, é importante **verificar a significância das variáveis no seu modelo**.

CAIXA DE DÚVIDAS

desafio conquer #08

- Melhorar o **modelo de precificação** de carros.
- **Prever o preço** dos imóveis de Melbourne, baseado nos **indicadores** disponibilizados.
- Baixe o arquivo **desafio_Conquer_08.xlsx** disponível para download no botão abaixo.

download

Conquer **notes**

quero++

site

KAGGLE
Find Open Datasets and Machine Learning Projects

livro

NOÇÕES DE PROBABILIDADE E ESTATÍSTICA
Marcos N. Magalhães e Antonio C. Pedroso de Lima

INTRODUÇÃO À ECONOMETRIA
Jeffrey M. Wooldridge