FIAP

MBA EM DATA SCIENCE & AI

APPLIED STATISTICS

AULA 5

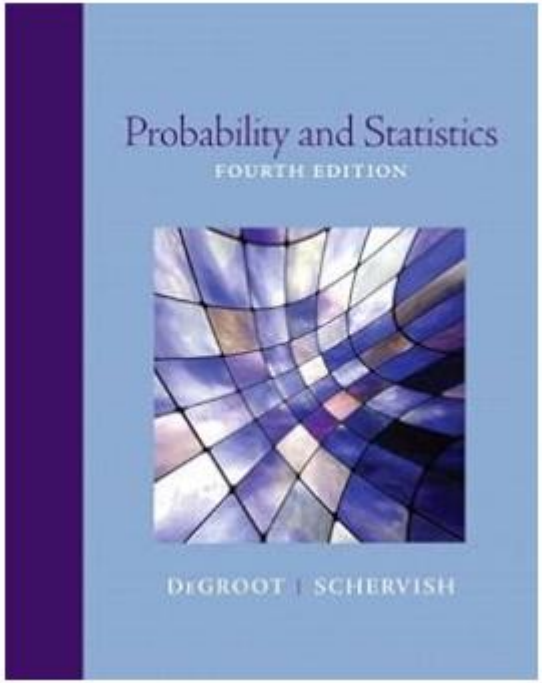
Teorema do Limite Central

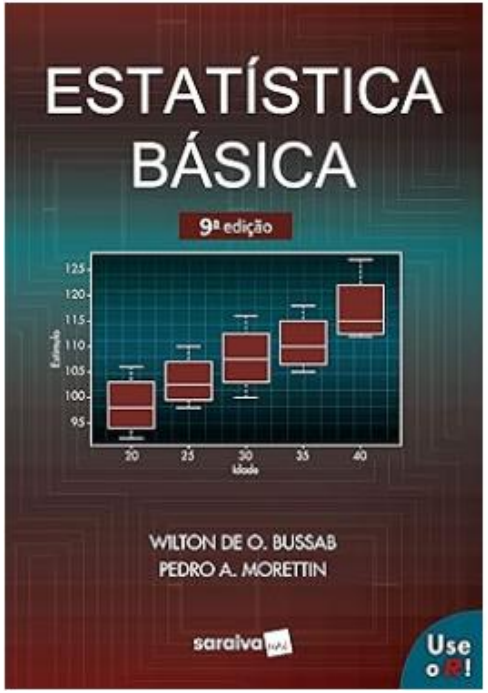
Intervalo de Confiança

Técnicas Supervisionadas

Regressão Linear Simples e Múltipla

Dicas de Leitura



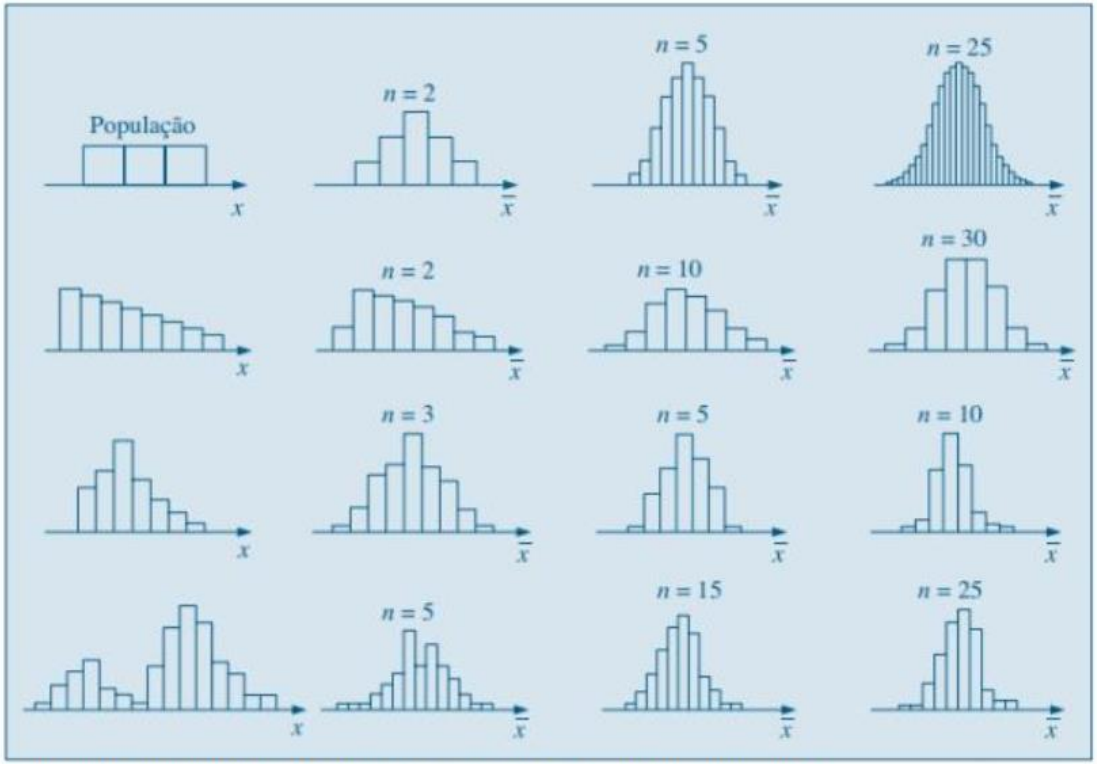


Teorema do Limite Central

Para amostras aleatórias simples (X_1, X_2, ..., X_n), retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\bar{X}$ aproxima-se, para *n* grande, de uma distribuição normal, com média μ e variância σ^2/n.

"Se o tamanho da amostra é suficientemente grande, a distribuição das médias amostrais pode ser aproximada por uma distribuição normal, mesmo que a população original não seja normalmente distribuída."

Teorema do Limite Central

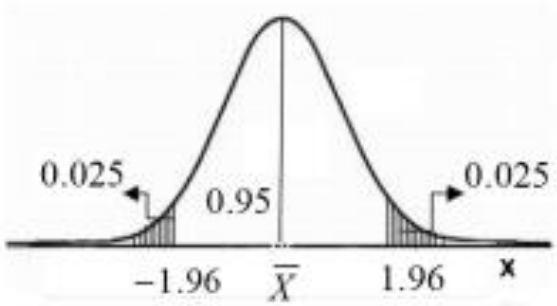


Fonte: Bussab & Morettin, Estatística Básica, 9ª edição, São Paulo: Saraiva, 2017.

Figura 1 - Histogramas correspondentes às distribuições amostrais de $\bar{X}$ para amostras extraídas de algumas populações.

Intervalos de Confiança

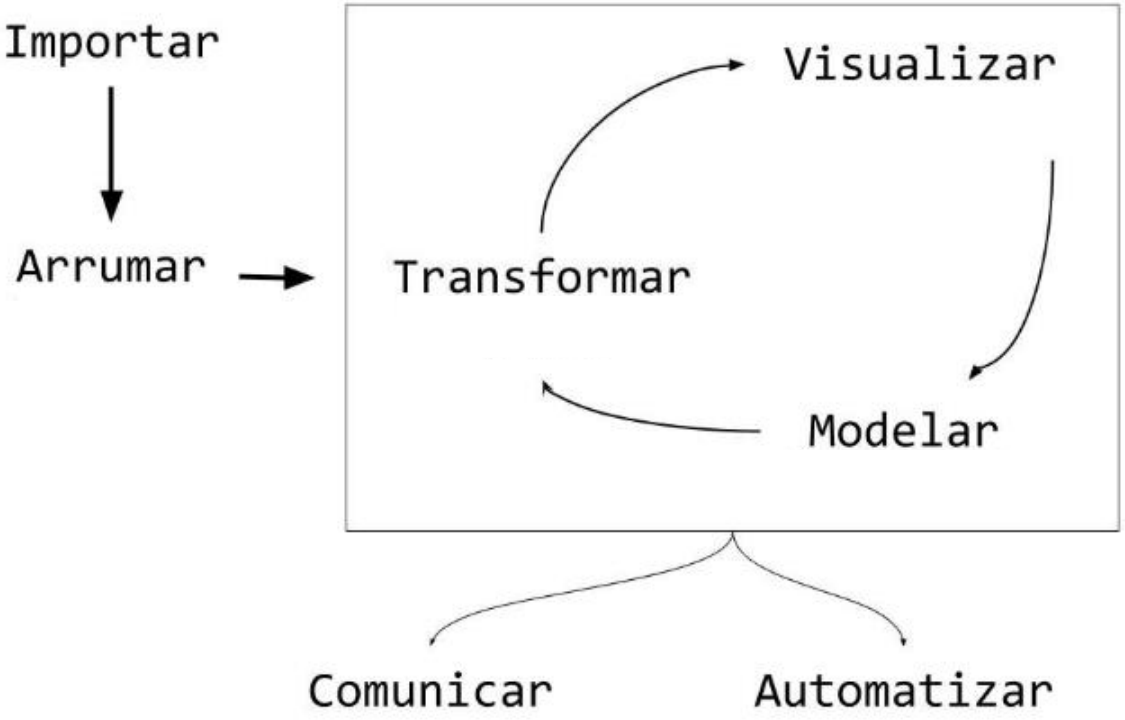
- Intervalos de confiança $P[(\bar{x} - 1.96.dp(\bar{x}) \leq \bar{X} \leq \bar{x} + 1.96.dp(\bar{x})] = 0.95$



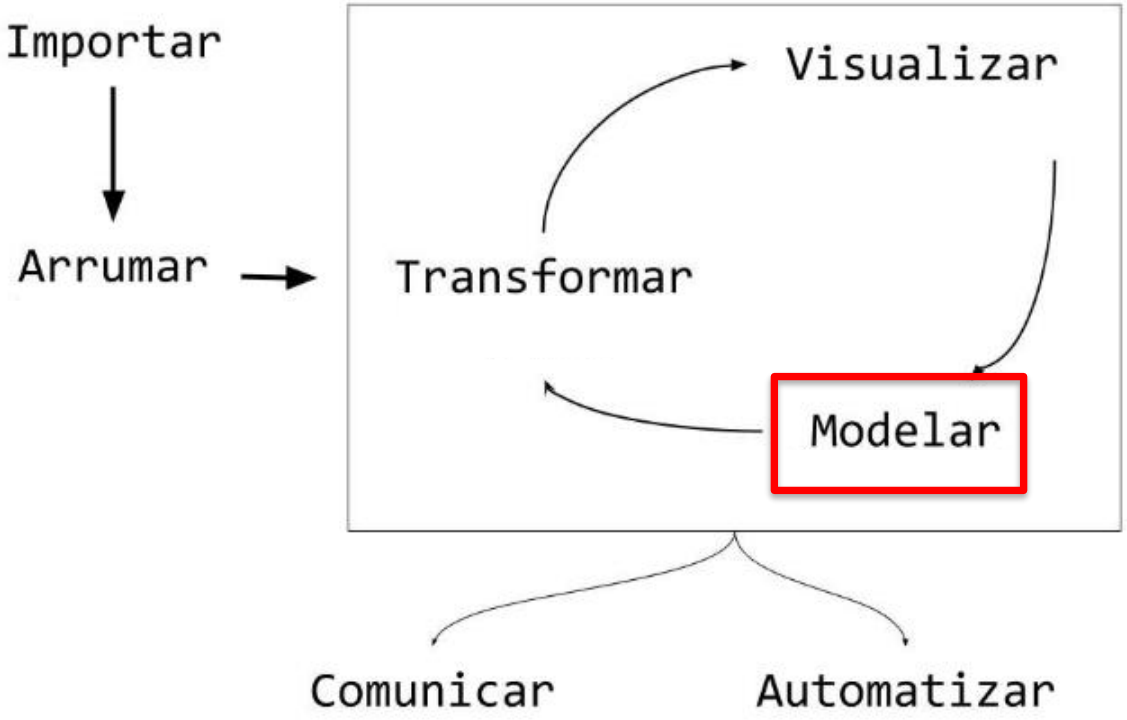
Distribuição da média amostral segundo o modelo normal com parâmetros ($\bar{x}; dp(\bar{x})$

O uso da distribuição normal como modelo para a distribuição da média amostral possibilita esperar que 95% das estimativas sejam diferentes do valor populacional por no máximo 1.96 desvios padrão.

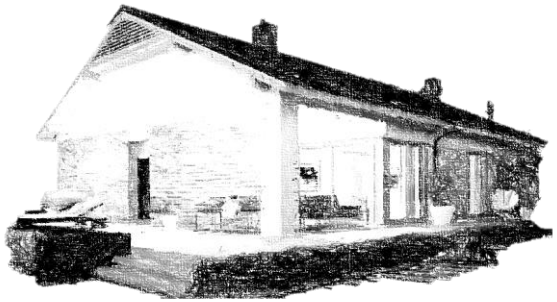
Ciência de dados (Data Science)



Ciência de dados (Data Science)

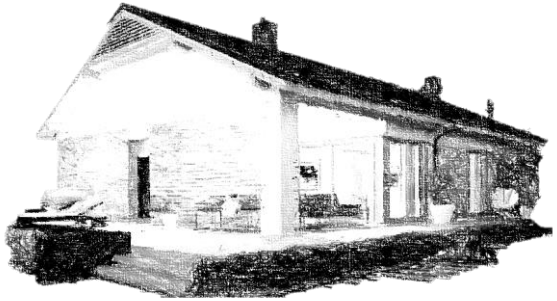


O que é um modelo?



Modelo 1

Posso ter mais de um modelo?



Modelo 1

Modelo 2

Um modelo é só um modelo!

Realidade

Modelos

Isso não é um cachimbo!

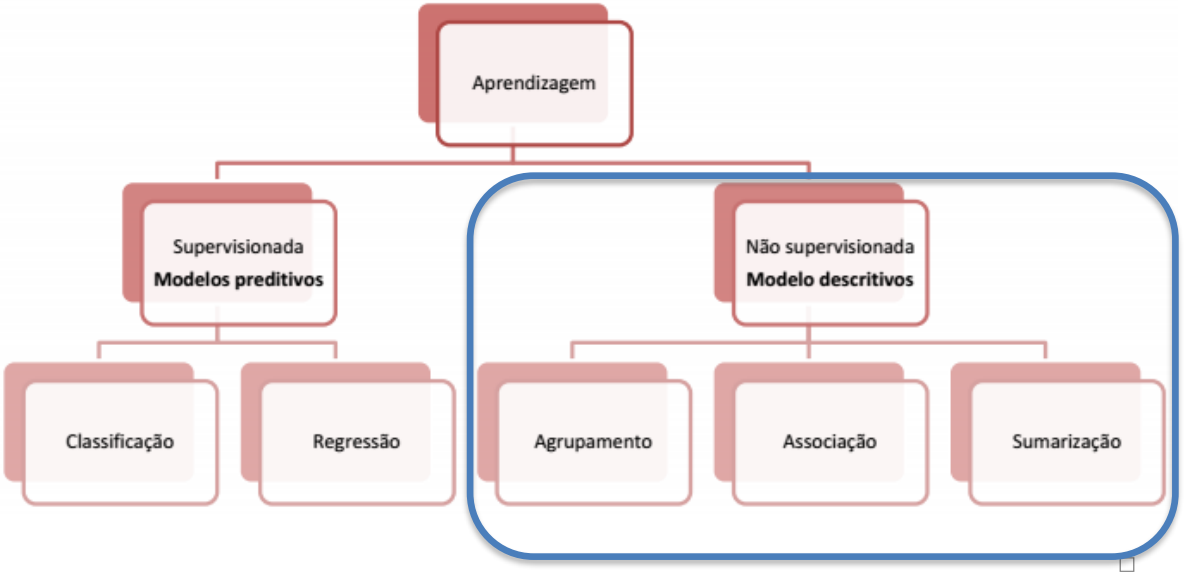
René Magritte, 1929

Aprendizagem de máquina

As técnicas de aprendizagem de máquina envolvem diversas finalidades, podendo ser supervisionadas ou não supervisionadas.

Aprendizagem não supervisionada

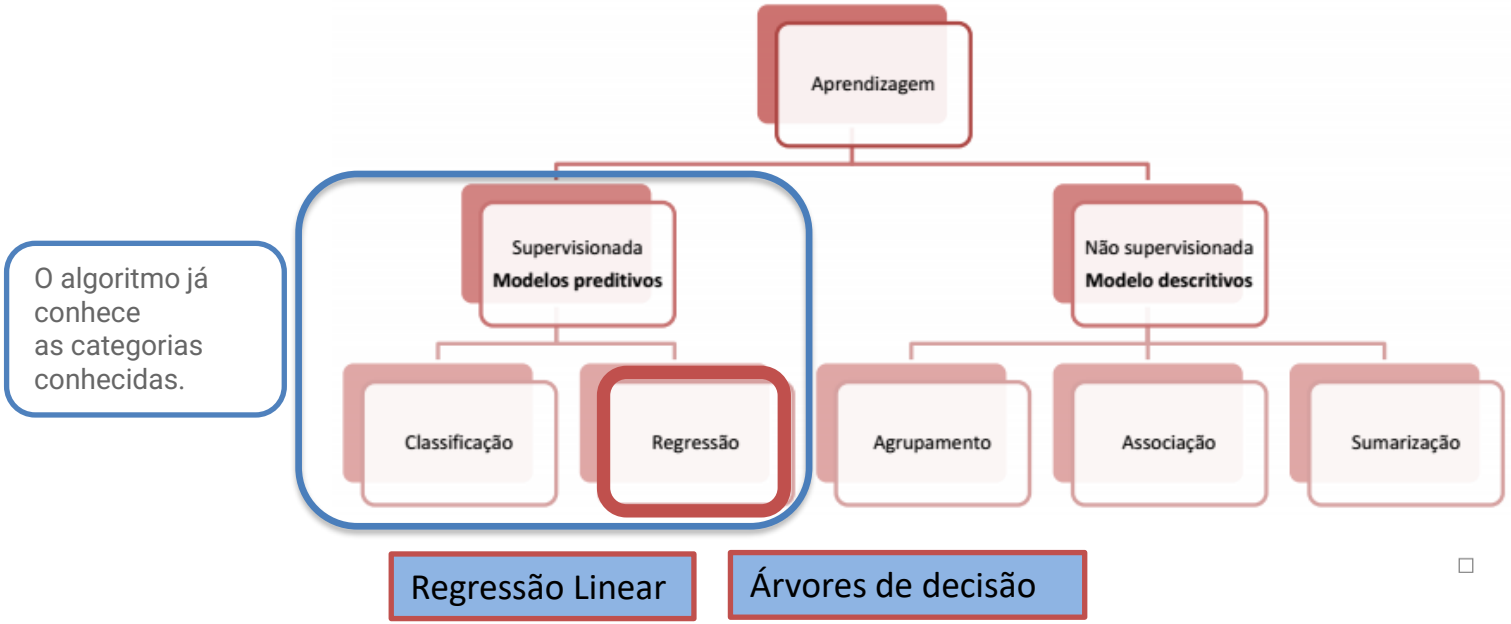
As técnicas de Machine Learning envolvem diversas finalidades, podendo ser supervisionadas ou não supervisionadas.



O algoritmo rotula através de padrões oriundo dos dados.

Aprendizagem supervisionada

As técnicas aprendizagem de máquina envolvem diversas finalidades, podendo ser supervisionadas ou não supervisionadas.

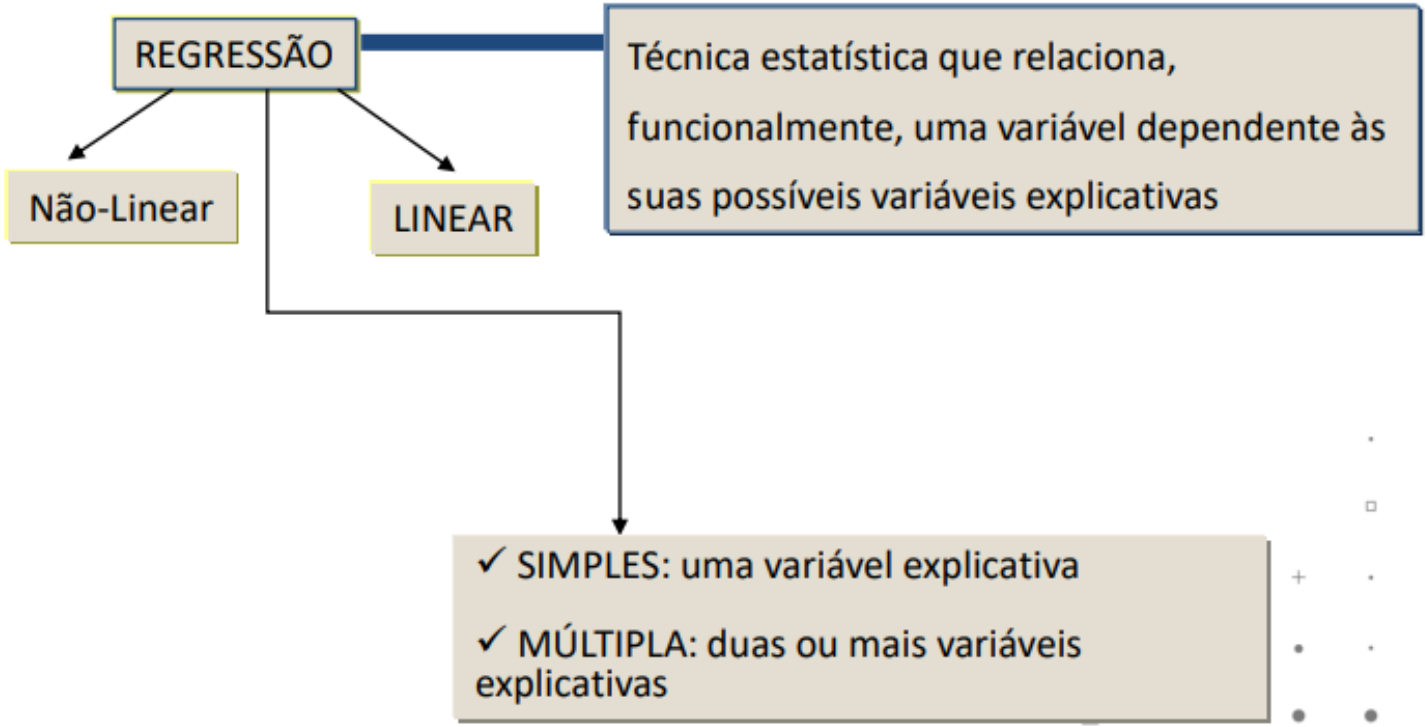


TÉCNICAS SUPERVISIONADAS

Regressão

- As técnicas quantitativas são aplicadas nas condições:
  - Informações do passado disponíveis;
  - Informações quantificáveis em forma numérica;
  - Assumir a hipótese de que algo dos padrões do passado irá se repetir no futuro (hipótese de continuidade).

Regressão


REGRESSÃO
Técnica estatística que relaciona, funcionalmente, uma variável dependente às suas possíveis variáveis explicativas
Não-Linear
LINEAR
✓ SIMPLES: uma variável explicativa
✓ MÚLTIPLA: duas ou mais variáveis explicativas

REGRESSÃO LINEAR SIMPLES

Exemplo 1

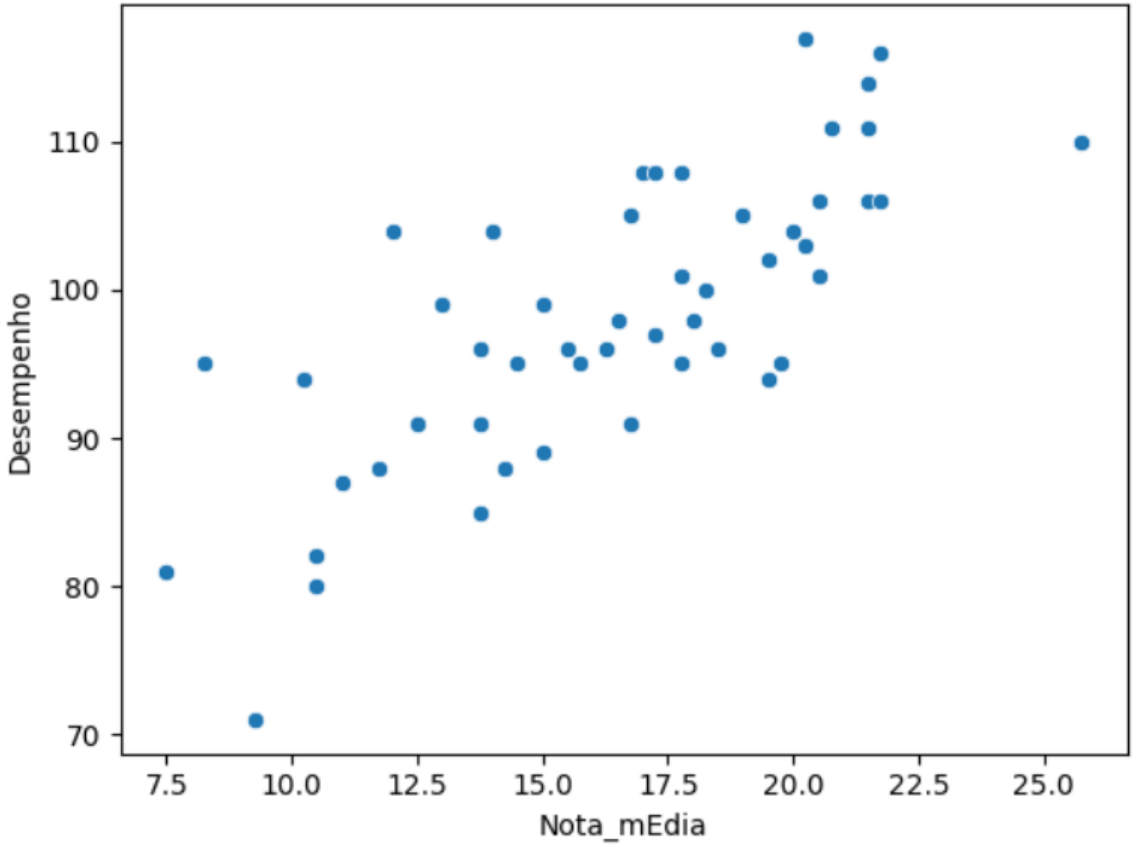
- O departamento de RH de uma empresa deseja avaliar a eficácia dos testes aplicados para a seleção de funcionários. Para tanto, foi sorteada uma amostra aleatória de 50 funcionários que fazem parte da empresa e que passaram pelo processo de seleção que utilizou os tais testes. Para cada um dos funcionários foi registrada a nota média nos testes de criatividade, raciocínio mecânico, raciocínio abstrato e habilidade matemática (notas de 0 a 26). Ainda, após 6 meses da contratação, foi calculado um escore que indica o seu desempenho profissional (0 a 120).

- Pergunta: existe alguma relação entre o escore de desempenho dos funcionários e a nota média nos testes?

Por onde começar ?

- **Diagrama de dispersão:** recurso gráfico que nos permite visualizar o comportamento conjunto das duas variáveis.
- **o Coeficiente de correlação linear:** mede a intensidade da associação linear existente entre as variáveis.

Diagrama de dispersão



Coeficiente de Correlação Linear

- Definição: Medida de associação linear entre duas variáveis quantitativas (varia entre –1 e +1).
  - **Valores próximos a +1:** indicam forte relação linear positiva;
  - **Valores próximos a -1:** indicam forte relação linear negativa;
  - **Valores próximos a zero:** indicam ausência de relação linear.

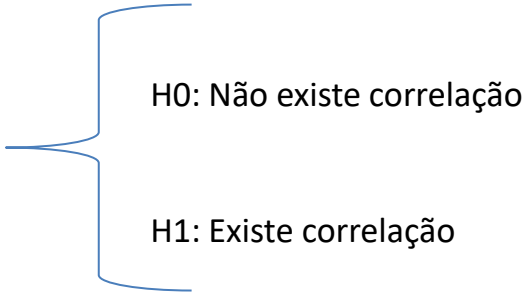
Calculando...

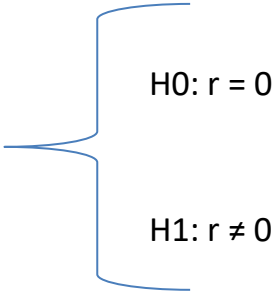
```
pearsonr(df['Nota_mEdia'], df['Desempenho'])
```

```
PearsonRResult(statistic=0.7621262240493802, pvalue=1.2828363756657115e-10)
```

A estatística exibida é dita **correlação de Pearson**

Correlação





$p < \alpha$: Rejeita a Hipótese Nula, ou seja, há correlação ao nível de significância α.

$p >= \alpha$: Não Rejeita a Hipótese Nula, ou seja, não há correlação ao nível de significância α.

Calculando...

```
pearsonr(df['Nota_mEdia'], df['Desempenho'])
```

```
PearsonRResult(statistic=0.7621262240493802, pvalue=1.2828363756657115e-10)
```

P-valor do teste de **correlação de Pearson**

há correlação ao nível de significância 5%.

Apesar de iniciar...

- Ainda não conseguimos **mensurar** a relação entre as medidas e muito menos **predizer** uma em relação a outra.
- Precisamos da técnica Regressão Linear

Regressão Linear Simples

Conceito

MODELO PROBABILÍSTICO

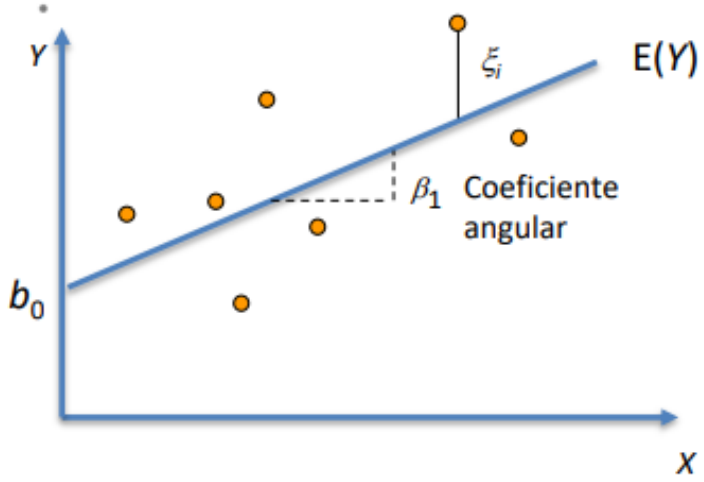
y = Componente Determinístico + Erro Aleatório

onde y é a variável dependente

Escrever a equação linear envolve dois parâmetros:

- ✓ O Intercepto de y
- ✓ A inclinação da reta

Regressão Linear Simples



$$E(Y) = \beta_0 + \beta_1 X$$

Inclinação populacional

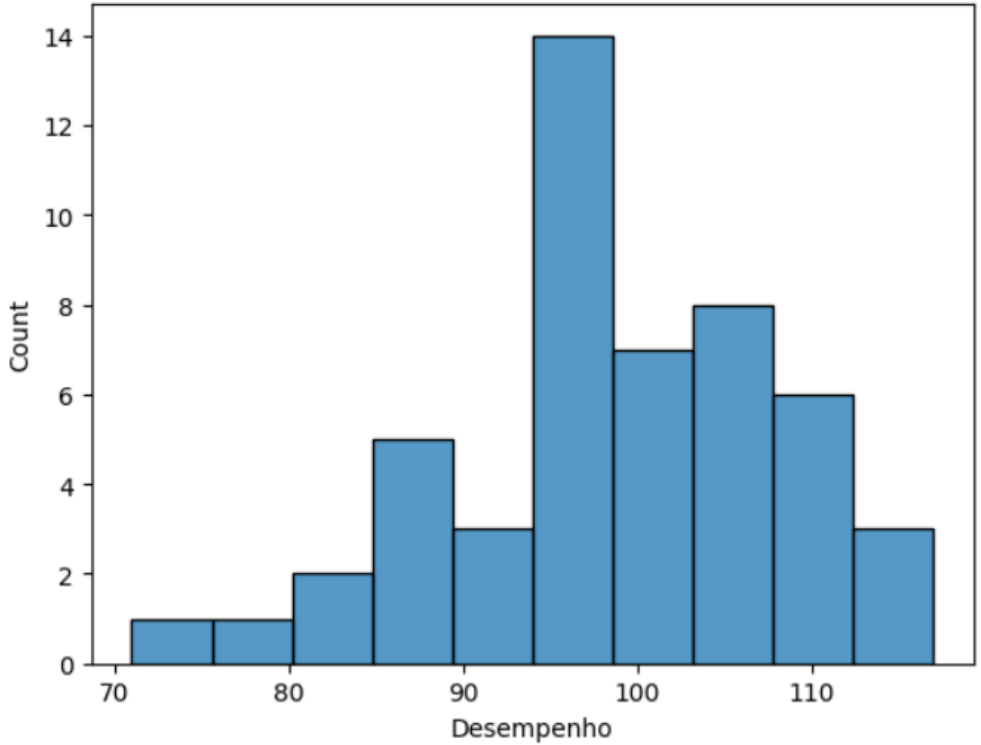
Intercepto populacional

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Voltando ao problema...

```
sbn.histplot(x = 'Desempenho', data = df, bins = 10)
```

```
<Axes: xlabel='Desempenho', ylabel='Count'>
```



Teste 1, 2, 3, ... Testando..

```
stats.shapiro(df['Desempenho'])
```

```
ShapiroResult(statistic=0.9837684570670189, pvalue=0.7176472808181578)
```

Testes de Normalidade

H0: Os dados seguem distribuição normal.

H1: Os dados não seguem distribuição normal.

Testes
- Shapiro-Wilk
- Anderson-Darling
- Kolmogorov-Smirnov

$p < \alpha$: Rejeita a Hipótese Nula, ou seja, não é normal ao nível de significância α.

$p >= \alpha$: Não rejeita a Hipótese Nula, ou seja, é normal ao nível de significância α.

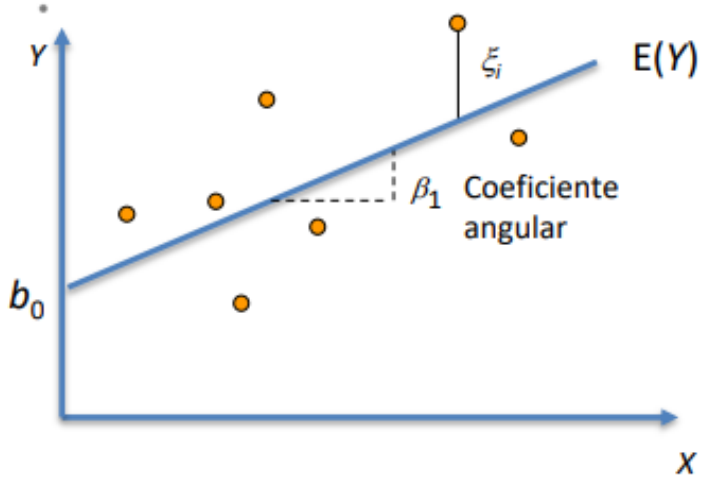
Teste de Normalidade de Shapiro Wilks

```
stats.shapiro(df['Desempenho'])
```

```
ShapiroResult(statistic=0.9837684570670189, pvalue=0.7176472808181578)
```

É considerado normal!

Regressão Linear Simples



$$E(Y) = \beta_0 + \beta_1 X$$

Inclinação populacional

Intercepto populacional

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Comando OLS da biblioteca `statsmodels`

```
import statsmodels.api as sm
```

```
results = sm.OLS(y, X_sm).fit()
```

Muito simples !!!

Lá no Python...

```
# mostrando as estatísticas do modelo
results.summary()
```

OLS Regression Results

Dep. Variable:	Desempenho	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.572
Method:	Least Squares	**F-statistic:**	66.51
Date:	Mon, 09 Sep 2024	**Prob (F-statistic):**	1.28e-10
Time:	19:04:00	**Log-Likelihood:**	-162.22
No. Observations:	50	**AIC:**	328.4
Df Residuals:	48	**BIC:**	332.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.5097	3.754	18.251	0.000	60.962	76.057
Nota_mEdia	1.8101	0.222	8.156	0.000	1.364	2.256

Omnibus:	0.813	**Durbin-Watson:**	2.255
Prob(Omnibus):	0.666	**Jarque-Bera (JB):**	0.906
Skew:	0.239	**Prob(JB):**	0.636
Kurtosis:	2.545	**Cond. No.**	71.1

Como interpretar os coeficientes?

Abuso de notação

$$\hat{y} = 68{,}51 + 1{,}81x$$

68,51: valor médio do desempenho dos funcionários que tiraram média igual a zero nos testes de admissão.

1,81: variação média no desempenho dos funcionários, quando aumenta-se a nota média obtida nos testes de admissão em 1 unidade.

Lá no Python...

```
# mostrando as estatísticas do modelo
results.summary()
```

OLS Regression Results

Dep. Variable:	Desempenho	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.572
Method:	Least Squares	**F-statistic:**	66.51
Date:	Mon, 09 Sep 2024	**Prob (F-statistic):**	1.28e-10
Time:	19:04:00	**Log-Likelihood:**	-162.22
No. Observations:	50	**AIC:**	328.4
Df Residuals:	48	**BIC:**	332.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.5097	3.754	18.251	0.000	60.962	76.057
Nota_mEdia	1.8101	0.222	8.156	0.000	1.364	2.256

Omnibus:	0.813	**Durbin-Watson:**	2.255
Prob(Omnibus):	0.666	**Jarque-Bera (JB):**	0.906
Skew:	0.239	**Prob(JB):**	0.636
Kurtosis:	2.545	**Cond. No.**	71.1

Teste de Hipóteses

TESTANDO OS PARÂMETROS B'S

$H_o: B_i = 0$

$H_1: B_i \neq 0$

$$t = \frac{B_i}{erro\_padrao(B_i)} \qquad \text{com gl = n - p}$$

Quando $t > t_{\alpha/2} \Rightarrow$ região de rejeição

$$IC : \bar{b}_i + t_{\alpha/2} Sb_i$$

Lá no Python...

```
# mostrando as estatísticas do modelo
results.summary()
```

OLS Regression Results

Dep. Variable:	Desempenho	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.572
Method:	Least Squares	**F-statistic:**	66.51
Date:	Mon, 09 Sep 2024	**Prob (F-statistic):**	1.28e-10
Time:	19:04:00	**Log-Likelihood:**	-162.22
No. Observations:	50	**AIC:**	328.4
Df Residuals:	48	**BIC:**	332.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.5097	3.754	18.251	0.000	60.962	76.057
Nota_mEdia	1.8101	0.222	8.156	0.000	1.364	2.256

Omnibus:	0.813	**Durbin-Watson:**	2.255
Prob(Omnibus):	0.666	**Jarque-Bera (JB):**	0.906
Skew:	0.239	**Prob(JB):**	0.636
Kurtosis:	2.545	**Cond. No.**	71.1

R^2 (R-quadrado)

O coeficiente de determinação ou R-quadrado representa a proporção da variância na variável dependente que é explicada pelo modelo de regressão linear.
É uma pontuação sem escala, ou seja, independentemente dos valores serem pequenos ou grandes,
o valor de R ao quadrado será menor que um.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

$\hat{y}$ – Valor predito
$\bar{y}$ – Valor médio

Lá no Python...

```
# mostrando as estatísticas do modelo
results.summary()
```

OLS Regression Results

Dep. Variable:	Desempenho	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.572
Method:	Least Squares	**F-statistic:**	66.51
Date:	Mon, 09 Sep 2024	**Prob (F-statistic):**	1.28e-10
Time:	19:04:00	**Log-Likelihood:**	-162.22
No. Observations:	50	**AIC:**	328.4
Df Residuals:	48	**BIC:**	332.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.5097	3.754	18.251	0.000	60.962	76.057
Nota_mEdia	1.8101	0.222	8.156	0.000	1.364	2.256

Omnibus:	0.813	**Durbin-Watson:**	2.255
Prob(Omnibus):	0.666	**Jarque-Bera (JB):**	0.906
Skew:	0.239	**Prob(JB):**	0.636
Kurtosis:	2.545	**Cond. No.**	71.1

Como ler o R^2 ?

- 58% das variações no desempenho dos funcionários após 3 meses de trabalho são explicadas pela nota média obtida nos testes de admissão.

Lá no Python...

```
# mostrando as estatísticas do modelo
results.summary()
```

OLS Regression Results

Dep. Variable:	Desempenho	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.572
Method:	Least Squares	**F-statistic:**	66.51
Date:	Mon, 09 Sep 2024	**Prob (F-statistic):**	1.28e-10
Time:	19:04:00	**Log-Likelihood:**	-162.22
No. Observations:	50	**AIC:**	328.4
Df Residuals:	48	**BIC:**	332.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.5097	3.754	18.251	0.000	60.962	76.057
Nota_mEdia	1.8101	0.222	8.156	0.000	1.364	2.256

Omnibus:	0.813	**Durbin-Watson:**	2.255
Prob(Omnibus):	0.666	**Jarque-Bera (JB):**	0.906
Skew:	0.239	**Prob(JB):**	0.636
Kurtosis:	2.545	**Cond. No.**	71.1

R^2 - Ajustado

O R quadrado ajustado é uma versão modificada do R quadrado, e é ajustado para o número de variáveis independentes no modelo, e sempre será menor ou igual a R^2.
Na fórmula abaixo de n é o número de observações nos dados e k é o número de variáveis independentes nos dados.

$$R^2_{adj} = 1 - \left[\frac{\left(1 - R^2\right)\left(n - 1\right)}{n - k - 1} \right]$$

Bom pra comparar modelos de regressões

Lá no Python...

```
# mostrando as estatísticas do modelo
results.summary()
```

OLS Regression Results

Dep. Variable:	Desempenho	**R-squared:**	0.581
Model:	OLS	**Adj. R-squared:**	0.572
Method:	Least Squares	**F-statistic:**	66.51
Date:	Mon, 09 Sep 2024	**Prob (F-statistic):**	1.28e-10
Time:	19:04:00	**Log-Likelihood:**	-162.22
No. Observations:	50	**AIC:**	328.4
Df Residuals:	48	**BIC:**	332.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.5097	3.754	18.251	0.000	60.962	76.057
Nota_mEdia	1.8101	0.222	8.156	0.000	1.364	2.256

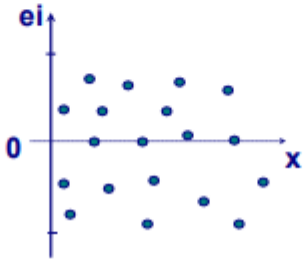
Omnibus:	0.813	**Durbin-Watson:**	2.255
Prob(Omnibus):	0.666	**Jarque-Bera (JB):**	0.906
Skew:	0.239	**Prob(JB):**	0.636
Kurtosis:	2.545	**Cond. No.**	71.1

Análise de Resíduos

Forma de avaliar se as suposições colocadas no desenvolvimento do modelo não foram violadas

$$\hat{e}_i = y_i - \hat{y}_i$$

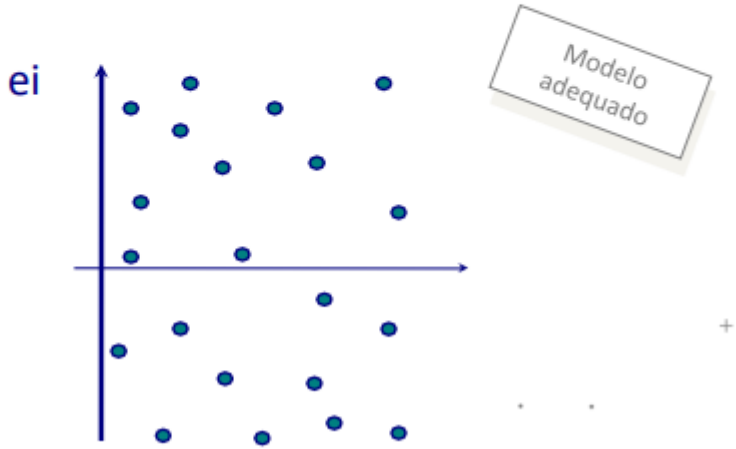
Pelo gráfico de dispersão, visualizamos o comportamento dos resíduos



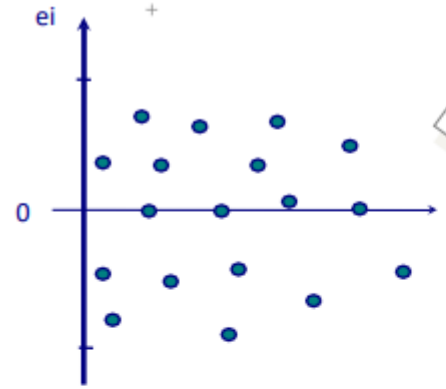
Análise de Resíduos

LINEARIDADE

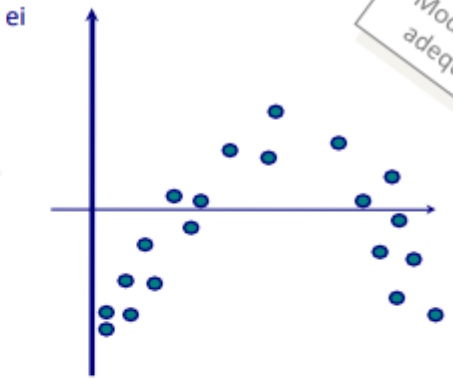
Gráfico de dispersão dos valores preditos (y) e o resíduo ⇨ os resíduos devem estar distribuídos aleatoriamente.



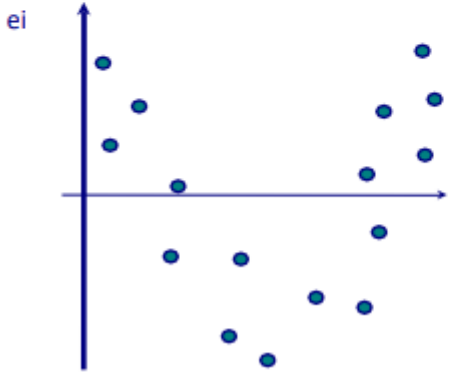
Análise de Resíduos



Modelo não adequado



Modelo não adequado



Modelo não adequado

o comportamento não aleatório dos resíduos podem ser eliminados ajustando no modelo um termo quadrático.

Análise de Resíduos

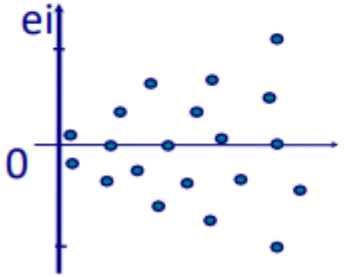
FIAP MBA+

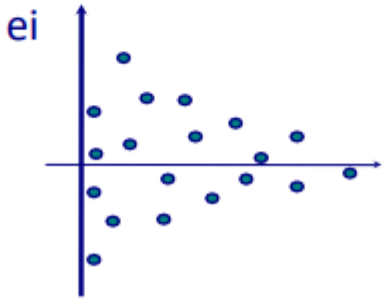
IGUALDADE DE VARIÂNCIA

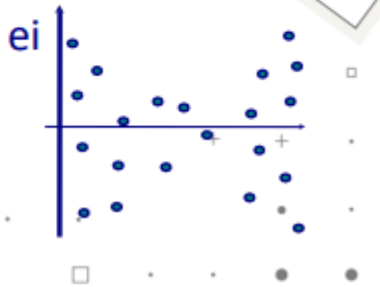
Quando o gráfico de dispersão dos Resíduos Studentizados, contra o valor predito, indica que a extensão dos resíduos aumentam com a magnitude dos valores preditos:

Então a suposição de igualdade da variância está violada

Modelo não adequado







Análise de Resíduos

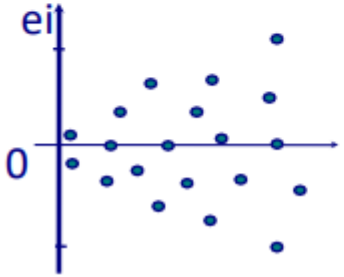
FIAP MBA+

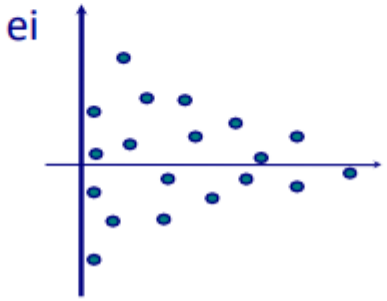
IGUALDADE DE VARIÂNCIA

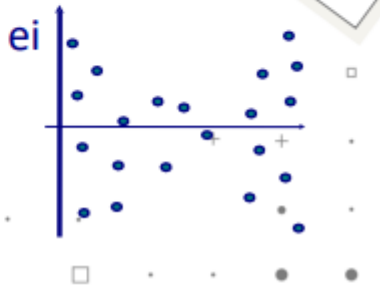
Quando o gráfico de dispersão dos Resíduos Studentizados, contra o valor predito, indica que a extensão dos resíduos aumentam com a magnitude dos valores preditos:

Então a suposição de igualdade da variância está violada

Modelo não adequado







Análise de Resíduos

NORMALIDADE

Pelo histograma dos resíduos padronizados pode-se analisar a suposição de normalidade.

Testar Normalidade com algum teste estudado (Shapiro- Wilks).

Perfomance de modelos

- R^2
- R^2 ajustado
- MAE
- MSE
- RMSE

MAE

O erro médio absoluto representa a média da diferença absoluta entre os valores reais e previstos no conjunto de dados.
Ele mede a média dos resíduos no conjunto de dados.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$\hat{y}$ – Valor predito
$\bar{y}$ – Valor médio

MSE

O erro quadrático médio representa a média da diferença quadrática entre os valores originais e previstos no conjunto de dados.
Ele mede a variância dos resíduos.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$\hat{y}$ – Valor predito
$\bar{y}$ – Valor médio

RMSE

Raiz quadrada do erro quadrático médio.
Mede o desvio padrão dos resíduos.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$\hat{y}$ – Valor predito
$\bar{y}$ – Valor médio

Como predizer o modelo?

FIAP MBA+

MAE

```
[46] MAE = mean_absolute_error(
         y_true = y, # valor verdadeiro
         y_pred = y_predito # predições
     )
     MAE.round(2)
```

```
4.96
```

MSE

```
[ ] MSE = mean_squared_error(
        y_true= y, # valor verdadeiro
        y_pred= y_predito # predições
    )
    MSE.round(2)
```

```
38.51
```

RMSE

```
RMSE = MSE**(1/2)
RMSE.round(2)
```

```
6.21
```

Regressão Linear Múltipla

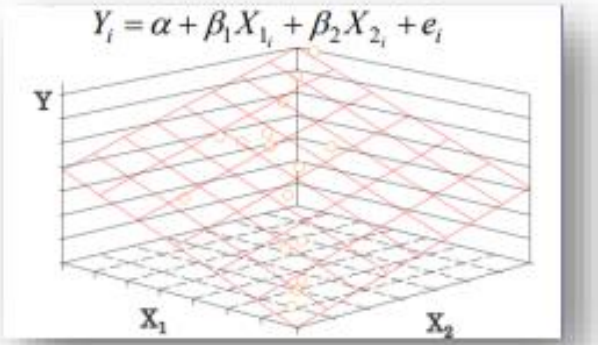
- Modelo Linear Múltiplo: $Y=B_o+B_1X_1+B_2X_2+B_3X_3+...+B_nX_n+ e$

$X_1,X_2,X_3,....,X_n$ = variáveis independentes

Y = variável dependente

B_o = constante

$B_1,B_2,B_3,....,B_n$ = coeficientes de regressão associados às n variáveis



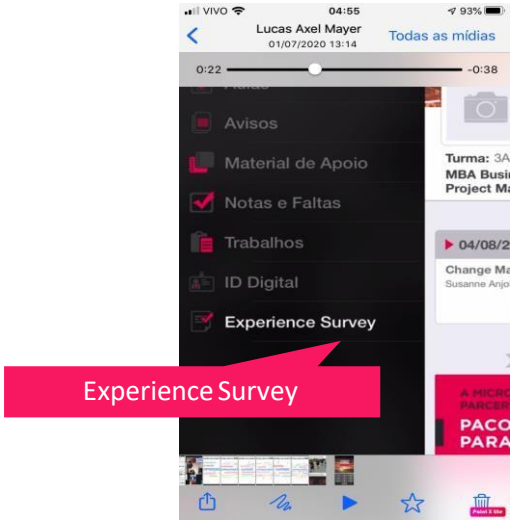
Exemplo 2

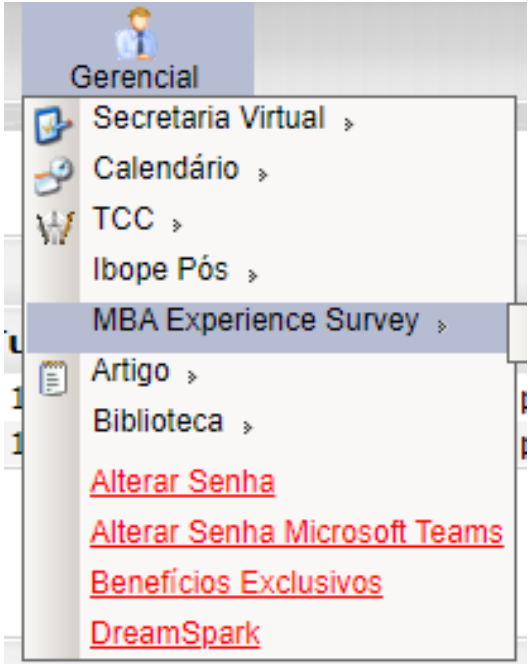
- Fazer um modelo de regressão linear para a base salario.csv e tentar predizer o salário.

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADO

in/lafphd

profleandro.ferreira@fiap.com.br

FIAP MBA+


FIAP