Regressão Linear Simples: $y = \beta_0 + \beta_1 \times x + \varepsilon$,

- y: variável dependente ou resposta
- β_0: interseção com o eixo das ordenadas
- β_1: declive da reta de regressão
- x: variável independente ou preditora
- ε: erro aleatórios em que $\varepsilon \sim N(0, \sigma^2)$

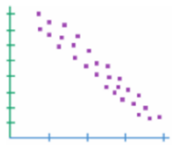
Resíduo: $e_i = y_i - \hat{y}_i$

Pressupostos: $\varepsilon \sim N(0, \sigma^2)$. O que significa que os resíduos devem:

- Ser independentes
- Seguir uma distribuição normal
- Ter média igual a zero
- Ter variância constante

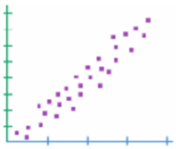
Coeficiente de Correlação (R): mede o grau da correlação ou associação linear entre duas variáveis quantitativas, em que $-1 \leq R \leq 1$.

- Se $R = -1$ é considerado como uma correlação negativa perfeita entre as duas variáveis (por exemplo, X e Y). O que significa que ao se aumentar em X vai haver uma diminuição na variável Y ou ao contrário.



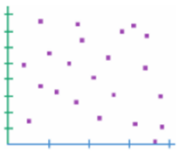
Correlação negativa forte

- Se $R = 1$ é considerado como uma correlação positiva perfeita entre duas variáveis (por exemplo, X e Y). O que significa que ao se aumentar em X vai haver, também, um aumento linear na variável Y ou ao contrário.



Correlação positiva forte

- Se $R = 0$, as duas variáveis não estão associadas linearmente. Mas poderá existir uma associação não linear.



Sem correlação

Coeficiente de determinação (R^2): avalia a proporção da variância da variável dependente que é explicada pelo modelo de regressão, em que $0 \leq R^2 \leq 1$. Quanto maior for o coeficiente de determinação, mais explicativo é o modelo.

Exercício: O Facebook é a rede social com mais utilizadores. Estima-se que sejam 2.9 bilhões de utilizadores ativos, mensalmente. Para avaliar o número total de interações foram registadas o número de partilhas e de comentários. Uma amostra aleatória foi extraída.

Analisar Output do SPSS:

Estatísticas Descritivas (Figura 1): apresenta o valor médio, desvio padrão e o número de observações para cada variável em estudo. Assim, através desta informação, sabe-se que o número total de interações é igual a 428.63, enquanto para o número de partilhas e de comentários são iguais a 47.48 e 17.43, respetivamente. Já o desvio padrão é igual a 1055.709, 124.337 e 58.629 para o total de interações, partilhas e comentários, respetivamente. Por fim, o número de observações, da amostra aleatória selecionada, é igual a 40.

Descriptive Statistics

	Mean	Std. Deviation	N
TotalInteractions	428.63	1,055.709	40
share	47.48	124.337	40
comment	17.43	58.629	40

Figura 1 - Estatísticas Descritivas.

Correlações (Figura 2): apresenta o valor da correlação para cada combinação de variáveis. Pelos resultados apresentados, sabe-se que a Correlação de Pearson entre Número Total de Interações e o Número de Partilhas é dado por: $\boldsymbol{R = 0.975}$. Existe uma associação linear positiva forte entre as duas variáveis, uma vez que a correlação é muito próxima de 1. Para verificar se a correlação é significativa é preciso testar as seguintes hipóteses:

$$H_0: \rho = 0 \text{ vs } H_1: \rho \neq 0$$

A decisão é tomada tendo em consideração o valor p, dado pelos valores que estão em Sig. (1-tailed). Assim, o **valor p < 0.001**, apesar de estar a apresentado como **0.000 ($\neq$ 0)** por o SPSS apresentar apenas as 3 casas decimais. Pelo que se pode decidir que a hipótese nula é rejeitada, para um nível de significância de 5%. O que significa que a correlação é significativa.

Para a comparação do Número Total de Interações com o Número de Comentários, $\boldsymbol{R = 0.961}$, o que significa que existe também uma associação linear positiva forte entre as duas variáveis. Também se pode concluir que a associação entre o Número Total de Interações e o Número de Comentários é significativa por o $\boldsymbol{valor\ p < 0.001} < 0.05$.

Correlations

		TotalInteractions	share	comment
Pearson Correlation	TotalInteractions	1.000	.975	.961
	share	.975	1.000	.990
	comment	.961	.990	1.000
Sig. (1-tailed)	TotalInteractions	.	.000	.000
	share	.000	.	.000
	comment	.000	.000	.
N	TotalInteractions	40	40	40
	share	40	40	40
	comment	40	40	40

Figura 2 - Correlações de Pearson.

Variáveis do modelo (Figura 3): Na presente tabela são apresentadas as variáveis que entraram no modelo, isto é, o número de comentários e de partilhas, bem como a variável dependente (número total de interações).

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	comment, share[b]	.	Enter

a. Dependent Variable: TotalInteractions

b. All requested variables entered.

Figura 3 - Variáveis do Modelo de Regressão Linear.

Resumo do modelo (Figura *4*): Nesta tabela são apresentadas as métricas que nos permitem concluir sobre a qualidade do modelo. O coeficiente de determinação, R^2, definido em R Square, é igual a 0.951. O que significa que 95.1% do número total de interações pode ser explicado pela variação da variável número de comentários e do número de partilhas.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.975[a]	.951	.949	239.357

a. Predictors: (Constant), comment, share

b. Dependent Variable: TotalInteractions

Figura 4 - Resumo do Modelo.

ANOVA (Figura *5*): Esta tabela é utilizada para testar se o modelo em estudo é estatisticamente significativo, em que as hipóteses a serem testadas são as seguintes:

H_0: O modelo, em estudo, não é significativo vs H_1: O modelo, em estudo, é significativo

A decisão é feita através do valor p, na coluna Sig. Assim, para este caso, tem-se que $valor\ p < 0.001 < 0.05$, pelo que se rejeita H_0, para o nível de significância de 5%. Logo, o modelo em estudo é estatisticamente significativo.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41346557.7	2	20673278.9	360.843	.000[b]
	Residual	2119787.67	37	57,291.559		
	Total	43466345.4	39			

a. Dependent Variable: TotalInteractions

b. Predictors: (Constant), comment, share

Figura 5 - ANOVA.

Coeficientes do modelo de regressão (Figura 6): Com esta tabela é possível definir a reta de regressão para o problema em questão, decidir se os valores de β_0, β_1e β_2 são estatisticamente diferentes de zero e apresenta o intervalo de confiança para cada um deles. Assim, através da primeira coluna (B) pode-se extrair os valores estimados para β_0, β_1e β_2. O que significa que $\hat{\beta}_0 = \mathbf{20.006}$, $\hat{\beta}_1 = \mathbf{9.819}$ e $\hat{\beta}_2 = \mathbf{-3.303}$. Logo, a reta de regressão linear, para prever o número total de interações, é dada por:

$$\hat{y} = 20.006 + 9.819\ \times Partilhas - 3.303 \times Comentários$$

Pode-se concluir que com o aumento de uma unidade do número de partilhas, existe um aumento médio no número total de interações, mantendo o número de comentários constante. Em relação ao aumento de uma unidade no número de comentários existe uma diminuição média do número de interações, mantendo o número de partilhas constantes.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	20.006	45.916		.436	.666	-73.028	113.041
	share	9.819	2.143	1.156	4.582	.000	5.477	14.162
	comment	-3.303	4.545	-.183	-.727	.472	-12.512	5.906

a. Dependent Variable: TotalInteractions

Figura 6 - Coeficientes do Modelo de Regressão.

Em termos de testes de hipótese, existem 3 análises a ter em consideração, em que a decisão deve de ser tomada a partir do valor p apresentado na coluna Sig. Assim, tem-se que:

- $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$, como o $\boldsymbol{valor\ p = 0.666} > 0.05$, pelo que não existem evidências estatísticas para se rejeita H_0. O que significa que a reta passa na origem.
- $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$, como o $\boldsymbol{valor\ p < 0.001} < 0.05$, então rejeita-se H_0. O que significa que o coeficiente para o número de partilhas é estatisticamente diferente de zero.
- $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$, como o $\boldsymbol{valor\ p = 0.472} > 0.05$, então não existem evidências estatísticas para se rejeitar H_0. Logo, não parece haver uma relação linear significativa entre o número total de interações e o número de comentários.

Outra informação que falta ser analisada são os intervalos de confiança, apresentados no final da tabela (*Lower Bound and Upper Bound*). Portanto, o intervalo de confiança de 95% para a interseção nos eixos das ordenadas é $\boldsymbol{]-73.028; 113.041[}$. Para o coeficiente do número de partilhas é $\boldsymbol{]5.477; 14.162[}$ e para o coeficiente do número de comentários é dado por $\boldsymbol{]-12.512; 5.906[}$.

Validação dos Pressupostos: Existem quatro pressupostos a serem validados para a aplicabilidade do modelo de regressão linear alcançado.

Para a verificação da média dos resíduos ser igual a zero, recorre-se às estatísticas descritivas dos resíduos, Figura 7. O intervalo de confiança de 95% para a média dos resíduos é de, aproximadamente, $\boldsymbol{]-0.492; 0.299[}$. Como o valor zero pertence ao intervalo de confiança significa que a média dos resíduos pode ser considerada como zero.

Descriptives

			Statistic	Std. Error
Studentized Residual	Mean		-.0966198	.19557077
	95% Confidence Interval for Mean	Lower Bound	-.4921990	
		Upper Bound	.2989595	
	5% Trimmed Mean		-.1515323	
	Median		-.1284554	
	Variance		1.530	
	Std. Deviation		1.23689814	
	Minimum		-4.66349	
	Maximum		4.56003	
	Range		9.22352	
	Interquartile Range		.32204	
	Skewness		.530	.374
	Kurtosis		9.842	.733

Figura 7 - Estatísticas Descritivas para os Resíduos.

O próximo pressuposto a ser validado é a normalidade dos resíduos. Para a validação deste pressuposto existem duas possibilidades:

- Teste da Normalidade de Kolmogorov-Smirnov, Figura 8. As primeiras três colunas (Statistic, df, sig.) são relativas ao teste de Kolmogorov-Smirnov, em que as hipóteses a serem testadas são as seguintes:

  $$H_0: \varepsilon \sim N(\mu, \sigma) \text{ vs } H_1: \varepsilon \nsim N(\mu, \sigma)$$

  A decisão é tomada a partir do valor presente em *Sig.* que representa o valor p. Como o $\boldsymbol{valor\ p < 0.001 < 0.05}$, rejeita-se H_0, considerando o nível de significância de 5%. Logo, os resíduos não seguem uma distribuição normal.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Studentized Residual	.305	40	.000	.656	40	.000

a. Lilliefors Significance Correction

Figura 8 - Teste da Normalidade dos Resíduos.

- Visualização gráfica do *Normal Q-Q Plot,* Figura 9. Quando a grande maioria dos valores estão dispostos segundo a diagonal significa que os resíduos seguem uma distribuição normal. No entanto, para o problema em questão isso não acontece pois existem valores que estão afastados da diagonal. Logo, os resíduos não seguem uma distribuição normal.

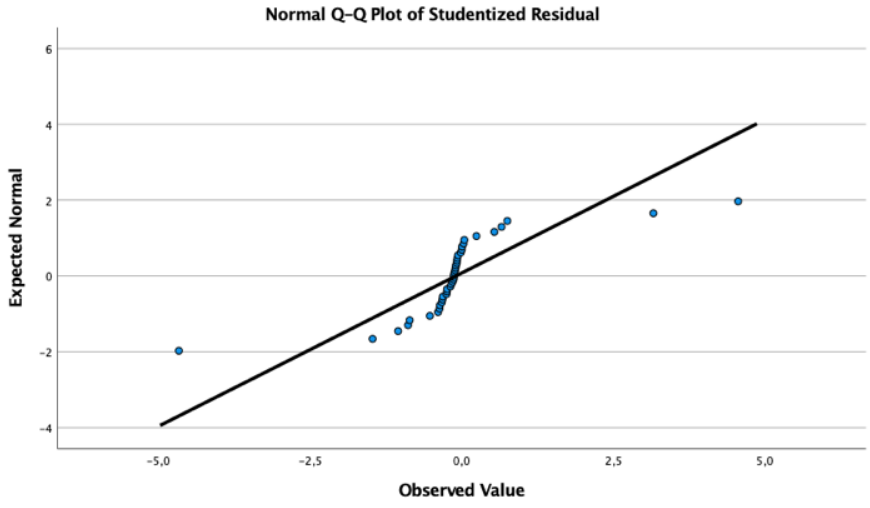


Figura 9 - Normal Q-Q Plot para os Resíduos.

Como existem valores muito afastados da diagonal pode ser pontos discrepantes, isto é, *outliers*. Para essa verificação recorre-se ao gráfico da caixa

de bigodes, Figura 10. Com esta visualização é possível verificar que existem vários *outliers*. Uma análise aprofundada deverá ser feita para perceber o porquê de serem pontos discrepantes.

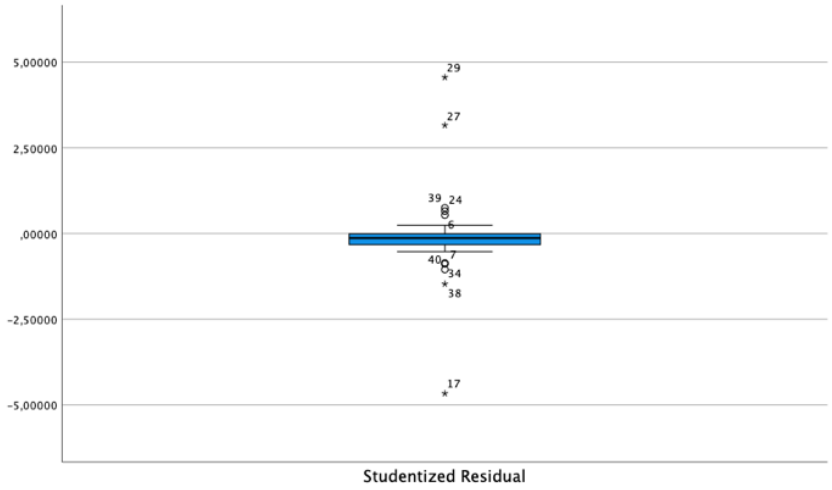


Figura 10 - Caixa de Bigodes para os Resíduos.

Por fim, falta verificar se a variância dos resíduos é constante e se são independentes. A partir do gráfico de resíduos, Figura 11, é possível visualizar que os valores estão dispostos aleatoriamente e não existe nenhum padrão. Logo, os pressupostos da variância dos resíduos ser constante e de os resíduos serem independentes são validados.

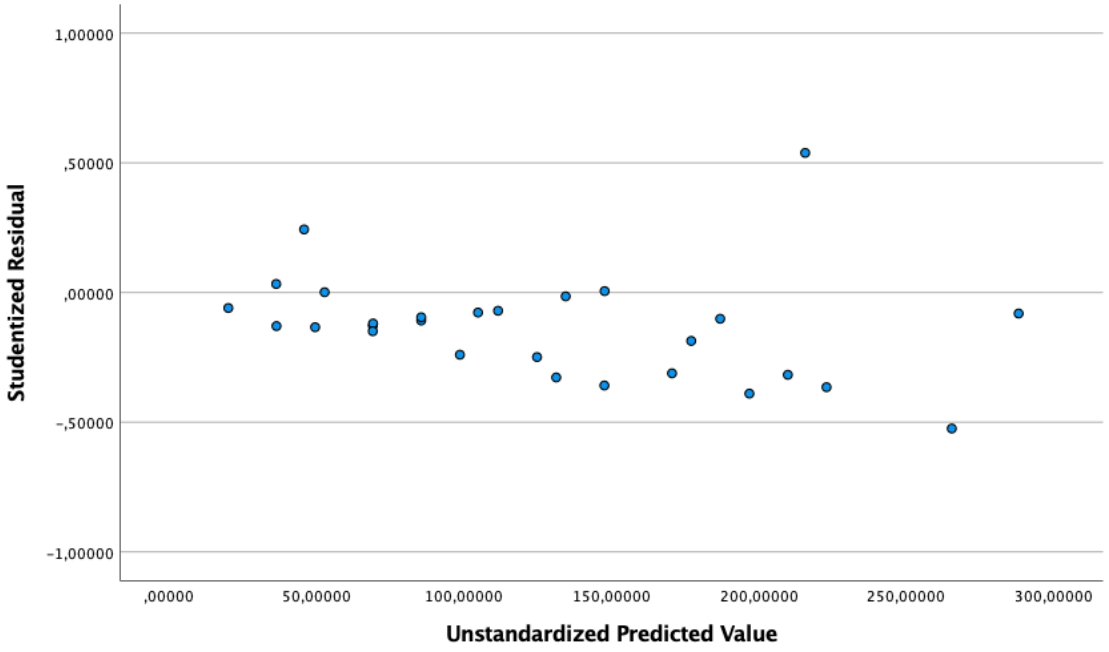


Figura 11 - Gráfico dos Resíduos.