Instituto Superior de Contabilidade e Administração de Coimbra
Instituto Politécnico de Coimbra

Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Trabalho N. º2 de
Análise Estatística de Dados

BÁRBARA XAVIER SOBRAL (2024104488)

MARIA RIBEIRO MARGARIDO (2024104121)

MARIANA SOFIA MENDES PRATA DE ALMEIDA (2024102807)

COIMBRA

10-11-2024

Parte I

Uma determinada companhia aérea pretende analisar a satisfação dos seus passageiros, mais concretamente, conhecer os atributos que podem determinar a sua satisfação. Foi aplicado um questionário de satisfação aos clientes, tendo-se obtido dados para uma amostra com dimensão 3695.

Para se avaliar o ponto mencionado anteriormente vamos ter como base o ficheiro 'Dados1_Trabalho2.sav' de onde se gerou uma amostra aleatória de dimensão N=3695 – 5 * 4 = 3675, obtida através do programa SPSS29 e seguindo os seguintes passos: Data > Select Cases > Random sample of cases > Sample > Exactly > 3675 … 3695.

A amostra em estudo apresenta 16 variáveis que podem ser divididas em dois grupos:

Variáveis quantitativas:

ID: número de identificação do cliente.

Variáveis qualitativas:

Tipo: tipo de viagem: 1 - viagem de negócios; 2 - viagem pessoal;

Classe: classes do avião: 1 - classe económica; 2 - classe executiva; 3 - primeira classe;

Avaliadas numa escala de 1 a 5 (1=muito insatisfeito e 5=muito satisfeito)

WIFI: grau de satisfação com o serviço de Wi-Fi a bordo;

CONV: grau de satisfação com a conveniência do horário de partida/chegada;

RES: grau de satisfação com a facilidade de reserva online;

POR: grau de satisfação com a localização do portão;

COM: grau de satisfação com a comida e bebida;

CHON: grau de satisfação com o check-in online;

ASS: grau de satisfação com o conforto do assento;

ENT: grau de satisfação com o entretenimento a bordo;

SERV: grau de satisfação com o serviço a bordo;

PERN: grau de satisfação com o espaço para as pernas;

BAG: grau de satisfação com o manuseio de bagagem;

CHSER: grau de satisfação com o serviço de check-in;

LIMP: grau de satisfação com a limpeza.

Diferenças significativas

ASS e *Tipo*

O caminho utilizado para obter os outputs da primeira questão foi: Analyze > Nonparametric tests > Independent Samples > selecionar Customize analysis em Objetive > selecionar ***ASS*** para Test Fields e ***Tipo*** para Groups > Em Settings > Customize Tests > escolher Mann-Whitney U > RUN.

Uma vez que a distribuição da variável ***ASS*** não é caracterizada por parâmetros - a distribuição não assume normalidade, possui uma escala ordinal (1 a 5) e a amostra é numerosa (3675) - optou-se por utilizar um **teste não paramétrico**, de modo a garantir a validade dos resultados e a obter uma interpretação mais fiel da realidade que queremos estudar.

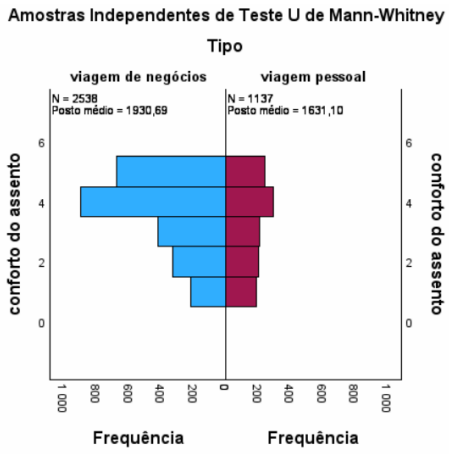
O **teste de Mann-Whitney U** foi escolhido uma vez que o objetivo é verificar se existem diferenças significativas entre as distribuições de satisfação com o conforto do assento (variável ***ASS***) entre **dois grupos independentes** em estudo: **viagem de negócios** e **viagem pessoal**.

Hipóteses:

H_0: Não há diferenças significativas entre as distribuições das avaliações dos assentos dos dois tipos de viagem.

H_1**:** Há diferenças significativas entre as distribuições das avaliações dos assentos dos dois tipos de viagem.

Gráfico 1 - Distribuição da variável ASS por Tipo de Viagem



Os gráficos de distribuição e boxplots ilustram que o conforto do assento é avaliado de

forma mais positiva nas viagens de negócios do que nas viagens pessoais, refletido pela diferença nos postos médios **(1930,69 contra 1631,10)**. Essa diferença sugere que passageiros de negócios tendem a atribuir uma avaliação mais alta ao conforto, possivelmente devido a expectativas diferentes associadas à natureza da viagem.

Neste sentido, para determinar se a diferença nos ranks médios é estatisticamente significativa, recorreu-se ao *valor-p*.

Testes não paramétricos

Sumarização de Teste de Hipótese

	Hipótese nula	Teste	Sig.[a,b]	Decisão
1	A distribuição de conforto do assento é igual nas categorias de Tipo.	Amostras Independentes de Teste U de Mann-Whitney	<,001	Rejeitar a hipótese nula.

a. O nível de significância é ,050.

b. A significância assintótica é exibida.

Tabela 1 - Sumarização de Teste de Hipótese

Amostras Independentes de Resumo de Teste U de Mann-Whitney

N total	3675
U de Mann-Whitney	1207606,000
Wilcoxon W	1854559,000
Estatística de teste	1207606,000
Erro padrão	28836,276
Estatística de Teste Padronizado	-8,158
Sinal assintótico (teste de dois lados)	<,001

Tabela 2 - Teste U de Mann-Whitney

Dado que o **valor-p < 0,001 é inferior a qualquer nível de significância usual** (neste caso, $\alpha = 0{,}05$), **rejeitamos a hipótese nula (H_0)**. Isto significa que há diferenças significativas entre as distribuições das avaliações dos assentos nos dois grupos, sugerindo que os clientes avaliam de forma significativamente diferente o conforto do assento nos dois tipos de viagem (negócios ou pessoal).

É importante notar que o tamanho das amostras é desigual: **2538 avaliações para viagens de negócios** e **1137 para viagens pessoais**. Esta discrepância (uma amostra tem mais do dobro da dimensão da outra) pode, em alguns casos, influenciar ligeiramente a precisão do teste, embora o teste de Mann-Whitney U seja robusto para tamanhos de amostra moderadamente diferentes (gráfico 2).

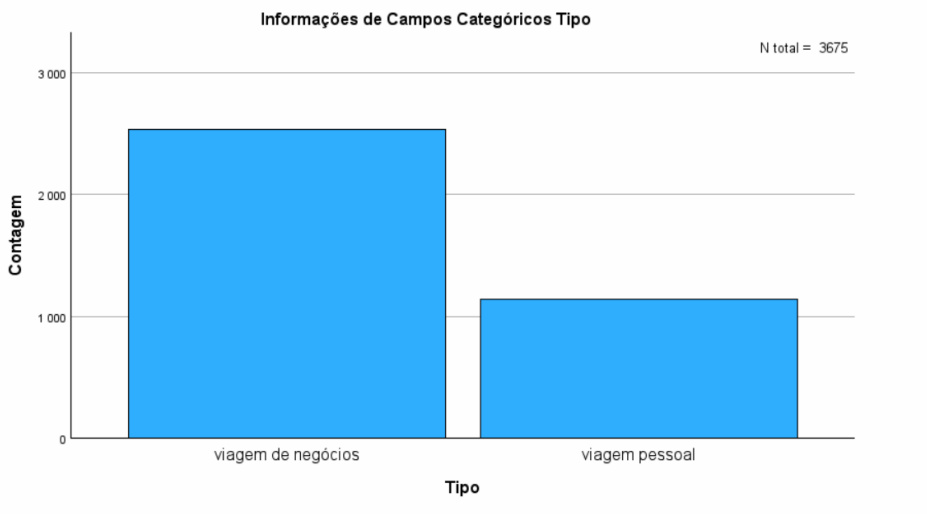


Gráfico 2 - Distribuição da variável Tipo

Os resultados indicam que existe uma diferença significativa na avaliação do conforto dos assentos entre viagens de negócios e viagens pessoais. Esta diferença sugere que o tipo de viagem pode influenciar a perceção dos passageiros sobre o conforto, o que pode ser útil para a companhia aérea ao planear melhorias nos serviços para cada grupo de clientes.

SERV* e *Classe

Inicialmente, foram gerados os outputs através do seguinte caminho: Analyze > Nonparametric tests > Independent Samples > selecionar Customize analysis em Objective > selecionar ***SERV*** para Test Fields e ***Classe*** para Groups > em Settings Customize Tests > Kruskal-Wallis e em Multiple comparisons utilizar Stepwise step-down > RUN.

De forma semelhante à questão anterior, procuramos verificar a ausência de diferenças entre as distribuições da variável, focando-nos, neste caso, no grau de satisfação com o serviço de bordo.

Neste sentido, foi utilizado o **teste de Kruskal-Wallis**, com vista a testar se existem diferenças significativas entre as distribuições da variável ***SERV*** nas três classes do avião: classe económica, classe executiva e primeira classe (mais de 2 grupos independentes).

Hipóteses:

H_0: Não há diferenças significativas entre as distribuições das avaliações do serviço de bordo nas três classes do avião.

H_1: Há diferenças significativas entre as distribuições das avaliações do serviço de bordo nas três classes do avião.

Sumarização de Teste de Hipótese

	Hipótese nula	Teste	Sig.[a,b]	Decisão
1	A distribuição de serviço a bordo é igual nas categorias de Classe.	Amostras Independentes de Teste de Kruskal-Wallis	<,001	Rejeitar a hipótese nula.

a. O nível de significância é ,050.

b. A significância assintótica é exibida.

Tabela 3 - Sumarização de Teste de Hipótese

Tabela 4 - Teste Kruskal-Wallis

Amostras Independentes de Resumo de Teste Kruskal-Wallis

N total	3675
Estatística de teste	152,830[a]
Grau de Liberdade	2
Sinal assintótico (teste de dois lados)	<,001

a. A estatística do teste está ajustada para empates.

Visto que o **valor-p < 0,001 é inferior a qualquer nível de significância usual** (neste caso, $\alpha = 0{,}05$), **rejeitamos a hipótese nula (H_0)**. Este resultado indica que há diferenças estatisticamente significativas entre as distribuições dos grupos, sugerindo que os clientes avaliam de forma distinta o serviço de bordo nas três classes do avião (económica, executiva e primeira classe).

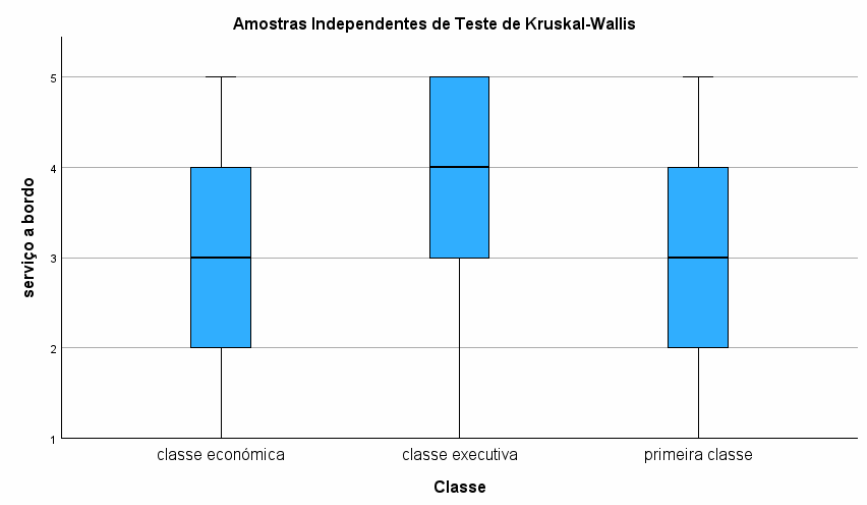


Gráfico 3 - Box Plot teste de Kruskal-Wallis variável Classe

No gráfico 3, observa-se que a mediana da satisfação com o serviço de bordo é mais elevada na classe executiva, em comparação com as outras classes.

Subconjuntos Homogêneos baseados em Classe

		Subconjunto 1	Subconjunto 2
Amostra[a]	classe económica	1630,197	
	primeira classe	1657,867	
	classe executiva		2054,969
Estatística de teste		,280	.[b]
Sig. (teste bilateral)		,596	.
Sig. Ajustada (teste bilateral)		,596	.

Subconjuntos homogêneos são baseados em significâncias assintóticas. O nível de significância é ,050.

a. Cada célula mostra o posto médio da amostra de Classe.

b. Não é possível calcular, pois o subconjunto contém somente uma amostra.

Tabela 5 - Teste Kruskal-Wallis

A tabela 5 permite-nos ainda constatar que o posto médio da **classe executiva** é estatisticamente diferente, em comparação com a classe económica e a primeira classe, comprovando que **o grau de satisfação com o serviço de bordo nessa classe é efetivamente superior ao das restantes**.

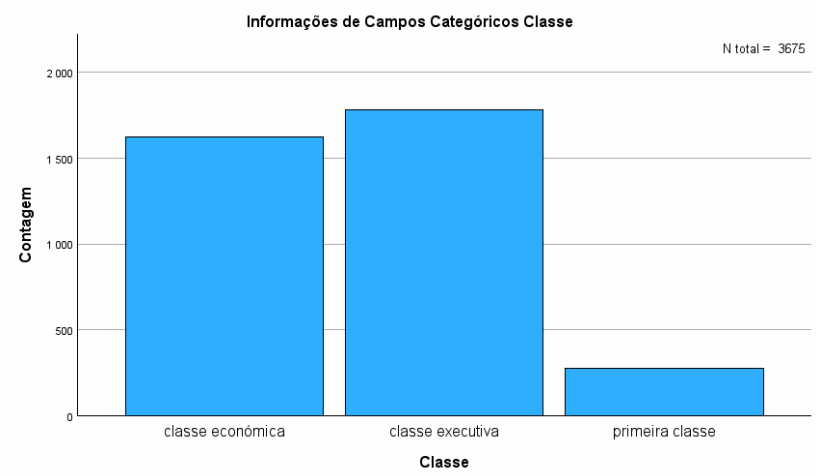


Gráfico 4 - Distribuição da variável Classe

Por fim, em termos de limitações na análise, é relevante considerar que o gráfico 4 destaca uma **desigualdade no tamanho das amostras** entre as diferentes classes, o que deve ser levado em conta ao interpretar os resultados do teste de Kruskal-Wallis. Embora o teste de Kruskal-Wallis possa ser usado, é importante considerar que:

- A **classe executiva** e a **classe económica** têm mais poder estatístico por terem tamanhos de amostra maiores.
- A **primeira classe** pode ter um efeito menos preciso ou menos confiável devido ao tamanho pequeno da amostra.

Análise Fatorial

Vamos agora realizar uma análise fatorial sobre as 13 variáveis, com o intuito de obter dimensões de satisfação mais gerais. Deste modo utilizámos a **estatística KMO** que varia entre 0 e 1 e realiza uma comparação das correlações simples com as parciais.

Os passos usados no SPSS29 para obter essa estatística foram: Analyze > Dimension Reduction > Factor > selecionar as variáveis ***WIFI***, ***CONV***, ***RES***, ***POR***, ***COM***, ***CHON***, ***ASS***, ***ENT***, ***SERV***, ***PERN***, ***BAG***, ***CHSER***, ***LIMP*** > em Descriptives selecionar Univariate descriptives e inicial solution em Statistics > selecionar Coefficients, Significance levels, Reproduced, KMO and Bartlett´s test em Correlation Matrix > em Extraction manter as opções usadas pelo Spss e selecionar Scree plot > em Rotation selecionar Varimax > e em Factor Scores selecionar Save as variables e Method Bartlett > OK.

Estatísticas Descritivas

	Média	Erro Desvio	Análise N
serviço de Wi-Fi a bordo	2,81	1,249	3675
conveniência do horário de partida/chegada	3,20	1,396	3675
facilidade de reserva online	2,88	1,309	3675
localização do portão	2,98	1,298	3675
comida e bebida	3,20	1,334	3675
check-in online	3,34	1,267	3675
conforto do assento	3,46	1,303	3675
entretenimento a bordo	3,39	1,328	3675
serviço a bordo	3,41	1,285	3675
espaço para as pernas	3,38	1,276	3675
manuseio de bagagem	3,64	1,163	3675
serviço de check-in	3,28	1,265	3675
limpeza	3,30	1,295	3675

Tabela 6 - Estatísticas Descritivas

Ao observar as estatísticas descritivas das 13 variáveis, podemos concluir que todas assumem valores muito idênticos relativamente à média e ao desvio padrão. Assim, a **média das variáveis** em análise **varia entre 2,81 e 3,64**, correspondendo ao serviço de

wifi a bordo e ao manuseio de bagagem, respetivamente. Por outro lado, o **desvio padrão oscila entre 1,163 e 1,396**, o que diz respeito ao manuseio de bagagem e à conveniência do horário de partida/chegada, respetivamente.

Matriz de correlações

		serviço de Wi-Fi a bordo	conveniência do horário de partida/chegada	facilidade de reserva online	localização do portão	comida e bebida	check-in online	conforto do assento	entretenimento a bordo	serviço a bordo	espaço para as pernas	manuseio de bagagem	serviço de check-in	limpeza
Correlação	serviço de Wi-Fi a bordo	1,000	,395	,685	,384	,171	,447	,157	,223	,123	,160	,109	,088	,160
	conveniência do horário de partida/chegada	,395	1,000	,527	,530	-,001	,077	,002	-,021	,097	-,002	,096	,127	,003
	facilidade de reserva online	,685	,527	1,000	,538	,055	,349	,048	,054	,052	,102	,038	,039	,028
	localização do portão	,384	,530	,538	1,000	,013	,008	,006	,022	-,002	-,022	,013	-,033	-,003
	comida e bebida	,171	-,001	,055	,013	1,000	,260	,581	,615	,069	,050	,034	,076	,646
	check-in online	,447	,077	,349	,008	,260	1,000	,447	,312	,171	,138	,094	,240	,355
	conforto do assento	,157	,002	,048	,006	,581	,447	1,000	,627	,157	,122	,083	,184	,685
	entretenimento a bordo	,223	-,021	,054	,022	,615	,312	,627	1,000	,439	,317	,372	,136	,691
	serviço a bordo	,123	,097	,052	-,002	,069	,171	,157	,439	1,000	,365	,528	,271	,134
	espaço para as pernas	,160	-,002	,102	-,022	,050	,138	,122	,317	,365	1,000	,381	,177	,109
	manuseio de bagagem	,109	,096	,038	,013	,034	,094	,083	,372	,528	,381	1,000	,251	,090
	serviço de check-in	,088	,127	,039	-,033	,076	,240	,184	,136	,271	,177	,251	1,000	,179
	limpeza	,160	,003	,028	-,003	,646	,355	,685	,691	,134	,109	,090	,179	1,000
Sig. (unilateral)	serviço de Wi-Fi a bordo		<,001	<,001	<,001	<,001	<,001	<,001	<,001	<,001	<,001	<,001	<,001	<,001
	conveniência do horário de partida/chegada	,000		,000	,000	,473	,000	,445	,105	,000	,448	,000	,000	,422
	facilidade de reserva online	,000	,000		,000	,000	,000	,002	,000	,001	,000	,011	,009	,046
	localização do portão	,000	,000	,000		,223	,303	,353	,088	,442	,096	,219	,023	,433
	comida e bebida	,000	,473	,000	,223		,000	,000	,000	,000	,001	,021	,000	,000
	check-in online	,000	,000	,000	,303	,000		,000	,000	,000	,000	,000	,000	,000
	conforto do assento	,000	,445	,002	,353	,000	,000		,000	,000	,000	,000	,000	,000
	entretenimento a bordo	,000	,105	,000	,088	,000	,000	,000		,000	,000	,000	,000	,000
	serviço a bordo	,000	,000	,001	,442	,000	,000	,000	,000		,000	,000	,000	,000
	espaço para as pernas	,000	,448	,000	,096	,001	,000	,000	,000	,000		,000	,000	,000
	manuseio de bagagem	,000	,000	,011	,219	,021	,000	,000	,000	,000	,000		,000	,000
	serviço de check-in	,000	,000	,009	,023	,000	,000	,000	,000	,000	,000	,000		,000
	limpeza	,000	,422	,046	,433	,000	,000	,000	,000	,000	,000	,000	,000	

Tabela 7 - Matriz de correlações

A análise da matriz de correlações assume uma elevada importância, uma vez que permite identificar grupos de variáveis que apresentam fortes correlações entre si, mas que, ao mesmo tempo, mantêm correlações relativamente baixas com outras variáveis. Destacam-se algumas correlações, nomeadamente a correlação de **0,685** entre o serviço de Wi-Fi a bordo e a facilidade de reserva online; o entretenimento a bordo apresenta correlações moderadas com o conforto do assento **(0,627)** e o check-in online **(0,615)**. Testando as seguintes as hipótese:

H_0: Matriz de correlações = Matriz identidade

H_1: Matriz de correlação ≠ Matriz identidade

Deste modo, deve ser testada a validade de aplicação deste tipo de análise, recorrendo ao teste de Bartlett e ao KMO.

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,770
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	19549,626
	gl	78
	Sig.	<,001

Tabela 8 - Teste de KMO e Bartlett

De acordo com o Teste de KMO e Bartlett, o **KMO** revela um valor de **0,770**. Este valor indica que a análise fatorial se caracteriza como **média**, visto que quanto mais próximo de 1, melhor é a adequação desta técnica.

No que diz ao **teste de Bartlett**, este revelou um valor de qui-quadrado de aproximadamente **19549,626**, o que demonstra ser altamente significativo.

De acordo com a análise deste teste, é possível concluir que **há uma correlação suficiente entre as variáveis**, o que justifica a aplicação da análise fatorial.

Deste modo, podemos concluir que devemos rejeitar H_0, visto que o valor-p (<0,01) tem um valor inferior a alfa (0.05), isto significa que de facto a matriz de correlações difere da matriz identidade.

Comunalidades

	Inicial	Extração
serviço de Wi-Fi a bordo	1,000	,699
conveniência do horário de partida/chegada	1,000	,613
facilidade de reserva online	1,000	,792
localização do portão	1,000	,720
comida e bebida	1,000	,726
check-in online	1,000	,771
conforto do assento	1,000	,733
entretenimento a bordo	1,000	,845
serviço a bordo	1,000	,666
espaço para as pernas	1,000	,460
manuseio de bagagem	1,000	,692
serviço de check-in	1,000	,472
limpeza	1,000	,783

Método de Extração: análise de Componente Principal.

Tabela 9 - Comunalidades

A comunalidade representa a proporção de variância de cada variável explicada pelas componentes. Posto isto, as variáveis "entretenimento a bordo" **(0,845)** , "facilidade de reserva online" **(0,792)** e "limpeza" **(0,783)** possuem comunalidades relativamente altas. Por outras palavras, isto significa que os componentes principais extraídos explicam uma grande parte da variabilidade dessas variáveis.

Por outro lado, embora as variáveis ‘espaço para as pernas’ (**0,460**) e ‘serviço de check-in’ **(0,472)** apresentem comunalidades baixas, indicando uma que a sua variabilidade não é explicada pelos componentes principais, a análise mantém valor ao identificar os fatores de maior influência na experiência do passageiro.

Variância total explicada

Componente	Autovalores iniciais			Somas de extração de carregamentos ao quadrado			Somas de rotação de carregamentos ao quadrado		
	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	3,719	28,610	28,610	3,719	28,610	28,610	2,968	22,829	22,829
2	2,449	18,840	47,450	2,449	18,840	47,450	2,530	19,458	42,287
3	1,785	13,731	61,182	1,785	13,731	61,182	2,095	16,116	58,402
4	1,017	7,823	69,005	1,017	7,823	69,005	1,378	10,603	69,005
5	,901	6,935	75,939						
6	,651	5,006	80,945						
7	,490	3,769	84,714						
8	,457	3,517	88,231						
9	,444	3,412	91,643						
10	,335	2,577	94,220						
11	,291	2,237	96,457						
12	,268	2,058	98,515						
13	,193	1,485	100,000						

Método de Extração: análise de Componente Principal.

Tabela 10 -Variância total explicada

Gráfico 5 -Variância total explicada

No seguimento da análise do output anterior, através da variância total explicada podemos reduzir a dimensão dos dados da amostra, transformando as variáveis em componentes principais. O primeiro componente explica cerca de **28%** da variabilidade dos dados, enquanto que em conjunto com o segundo, explicam **47,45%** da variância explicada.

Deste modo, foram selecionados os primeiros quatro componentes, visto que em conjunto são responsáveis por explicar **69%** do total, o que significa que os primeiros quatro componentes são fulcrais para entender a estrutura dos dados e facilitar a interpretação.

Outra forma para escolher o número das componentes é representar graficamente a percentagem de variância explicada por cada componente. De acordo com Scree Plot, à medida que esta porcentagem se reduz e a curva passa a ser quase paralela ao eixo das abcissas, como acontece a partir do componente 4, devem ser excluídas as componentes correspondentes. O gráfico 5, mostra esta informação de forma mais simplificada,

fazendo uma relação entre a componente e o seu autovalor, mostrando de forma notória as percentagens de explicação da variância.

Matriz de componente[a]

	Componente			
	1	2	3	4
serviço de Wi-Fi a bordo	,528	,616	-,081	-,186
conveniência do horário de partida/chegada	,251	,715	,014	,198
facilidade de reserva online	,385	,792	-,097	-,089
localização do portão	,206	,712	-,126	,393
comida e bebida	,649	-,296	-,418	,205
check-in online	,609	,099	-,129	-,611
conforto do assento	,730	-,311	-,319	-,025
entretenimento a bordo	,816	-,323	,015	,272
serviço a bordo	,470	-,072	,653	,116
espaço para as pernas	,386	-,045	,556	,014
manuseio de bagagem	,402	-,045	,703	,186
serviço de check-in	,351	-,033	,339	-,482
limpeza	,735	-,342	-,340	,097

Método de Extração: análise de Componente Principal.

a. 4 componentes extraídos.

Tabela 11 -Matriz de componente

Analisando a matriz de componente, conclui-se que foram extraídos apenas 4 componentes. O **componente 1** possui uma forte associação com variáveis, tais como "entretenimento a bordo" **(0.816)**, "conforto do assento" **(0.730)**, e "limpeza" **(0.735).** Isso sugere que este componente está fortemente relacionado à experiência a bordo e ao conforto dos passageiros. O **Componente 2** está associado a variáveis como "facilidade de reserva online" (**0.792**) e "conveniência do horário de partida/chegada" **(0.715)**, indicando um foco na conveniência e na eficiência dos serviços oferecidos.

O **Componente 3** apresenta uma maior relação com "manuseio de bagagem" (**0.703**) e "espaço para as pernas" **(0.556)**. Por fim, o **Componente 4** tem um associação notável para o "check-in online" (**-0.611**), o que pode indicar uma relação negativa entre o uso de check-in online e outros aspectos do serviço a bordo.

Ainda que sejam apenas considerados como pesos significativos, as variáveis com um peso acima de 0.5, a variável “Serviço de check in” apresenta pesos idênticos em 2 fatores. Então, vamos aplicar o método de rotação para tentar resolver estes problemas de interpretação.

Correlações reproduzidas

		serviço de Wi-Fi a bordo	conveniência do horário de partida/chegada	facilidade de reserva online	localização do portão	comida e bebida	check-in online	conforto do assento	entretenimento a bordo	serviço a bordo	espaço para as pernas	manuseio de bagagem	serviço de check-in	limpeza
Correlação reproduzida	serviço de Wi-Fi a bordo	,699[a]	,534	,715	,484	,156	,507	,224	,180	,129	,129	,093	,228	,187
	conveniência do horário de partida/chegada	,534	,613[a]	,643	,637	-,014	,101	-,048	,028	,098	,075	,115	-,026	-,046
	facilidade de reserva online	,715	,643	,792[a]	,620	,038	,380	,068	,032	,050	,057	,033	,119	,036
	localização do portão	,484	,637	,620	,720[a]	,056	-,028	-,040	,043	,009	-,017	,035	-,183	-,011
	comida e bebida	,156	-,014	,038	,056	,726[a]	,295	,695	,675	,077	,034	,019	-,003	,741
	check-in online	,507	,101	,380	-,028	,295	,771[a]	,470	,297	,124	,151	,036	,462	,399
	conforto do assento	,224	-,048	,068	-,040	,695	,470	,733[a]	,685	,154	,118	,079	,170	,750
	entretenimento a bordo	,180	,028	,032	,043	,675	,297	,685	,845[a]	,449	,342	,404	,171	,732
	serviço a bordo	,129	,098	,050	,009	,077	,124	,154	,449	,666[a]	,549	,673	,333	,159
	espaço para as pernas	,129	,075	,057	-,017	,034	,151	,118	,342	,549	,460[a]	,550	,319	,111
	manuseio de bagagem	,093	,115	,033	,035	,019	,036	,079	,404	,673	,550	,692[a]	,291	,090
	serviço de check-in	,228	-,026	,119	-,183	-,003	,462	,170	,171	,333	,319	,291	,472[a]	,107
	limpeza	,187	-,046	,036	-,011	,741	,399	,750	,732	,159	,111	,090	,107	,783[a]
Resíduo[b]	serviço de Wi-Fi a bordo		-,139	-,029	-,100	,015	-,059	-,067	,043	-,006	,032	,017	-,139	-,027
	conveniência do horário de partida/chegada	-,139		-,116	-,107	,013	-,024	,051	-,049	-,002	-,077	-,019	,153	,049
	facilidade de reserva online	-,029	-,116		-,082	,017	-,031	-,020	,022	,003	,045	,004	-,080	-,009
	localização do portão	-,100	-,107	-,082		-,044	,036	,046	-,021	-,012	-,004	-,022	,150	,008
	comida e bebida	,015	,013	,017	-,044		-,035	-,114	-,060	-,009	,015	,015	,079	-,094
	check-in online	-,059	-,024	-,031	,036	-,035		-,024	,015	,047	-,013	,058	-,222	-,044
	conforto do assento	-,067	,051	-,020	,046	-,114	-,024		-,058	,003	,003	,005	,013	-,065
	entretenimento a bordo	,043	-,049	,022	-,021	-,060	,015	-,058		-,009	-,025	-,032	-,035	-,041
	serviço a bordo	-,006	-,002	,003	-,012	-,009	,047	,003	-,009		-,184	-,144	-,062	-,025
	espaço para as pernas	,032	-,077	,045	-,004	,015	-,013	,003	-,025	-,184		-,169	-,142	-,003
	manuseio de bagagem	,017	-,019	,004	-,022	,015	,058	,005	-,032	-,144	-,169		-,040	,000
	serviço de check-in	-,139	,153	-,080	,150	,079	-,222	,013	-,035	-,062	-,142	-,040		,072
	limpeza	-,027	,049	-,009	,008	-,094	-,044	-,065	-,041	-,025	-,003	,000	,072	

Método de Extração: análise de Componente Principal.

a. Comunalidades reproduzidas

b. Os resíduos são computados entre as correlações observadas e reproduzidas. Há 27 (34,0%) resíduos não redundantes com valores absolutos maiores que 0,05.

Tabela 12 - Correlações reproduzidas

Os resíduos ajudam a avaliar a qualidade do modelo de componentes principais, destacando áreas onde o modelo poderia ser mais preciso na reprodução das correlações entre variáveis. A matriz dos resíduos resulta da diferença entre as matrizes das correlações observadas e das correlações estimadas pelo modelo. Considera-se que mais de 50% de resíduos inferiores a 0,05 é indicador de um bom ajustamento.

Isto significa que temos **66%** (100%-34%) de resíduos com valores absolutos inferiores a 0,05, ou seja, isto indica que temos um bom ajustamento. Em termos estatísticos, há uma quantidade significativa de variação não explicada pelo modelo, sugerindo a necessidade de melhoria ou ajuste no modelo de componentes principais usado.

Matriz de componente rotativa[a]

	Componente			
	1	2	3	4
serviço de Wi-Fi a bordo	,151	,713	,058	,405
conveniência do horário de partida/chegada	-,051	,774	,096	-,038
facilidade de reserva online	,003	,846	-,012	,276
localização do portão	,026	,812	-,006	-,243
comida e bebida	,850	,042	-,029	-,001
check-in online	,339	,191	,007	,788
conforto do assento	,819	-,002	,046	,245
entretenimento a bordo	,807	,036	,438	,017
serviço a bordo	,117	,040	,802	,090
espaço para as pernas	,062	,027	,658	,153
manuseio de bagagem	,048	,049	,829	,006
serviço de check-in	,012	-,047	,349	,590
limpeza	,873	-,002	,057	,131

Método de Extração: análise de Componente Principal.
Método de Rotação: Varimax com Normalização de Kaiser.[a]

a. Rotação convergida em 5 iterações.

Tabela 13 - Matriz de componente rotativa

De acordo com critério de Kaiser, os componentes com autovalores superiores a 1, são considerados relevantes. Se analisarmos os autovalores dos componentes, veremos que

os quatro primeiros componentes têm autovalores significativos, **3,719, 2,449, 1,785 e 1,017**, respetivamente, como podemos visualizar na **tabela 10**.

Esta matriz de rotação tem como objetivo principal definir um novo conjunto de pesos das variáveis para cada componente. O método mais eficaz, VARIMAX, pretende que, para cada componente existam apenas alguns pesos significativos, ou seja, quanto mais próximo de 1, mais forte se define a associação entre a variável e a componente. Posto isto:

Fator 1: comida e bebida; conforto do assento; entretenimento a bordo; limpeza
Fator 2: serviço de wifi a bordo; conveniência do horário de partida e chegada; facilidade de reserva online; localização do portão
Fator 3: serviço a bordo; espaço para as pernas; manuseio de bagagem
Fator 4: check-in online; serviço de check-in

Parte II

Pretende-se avaliar alguns determinantes do consumo de eletricidade dos consumidores residenciais de um certo país. Obteve-se informação relativamente a várias variáveis, que constam no ficheiro 'Dados2_Trabalho2.sav'. A partir desse ficheiro gerou-se uma amostra aleatória de dimensão N=885 – 5 * 4 = 865, que foi obtida através do programa SPSS29 e através dos seguintes passos: Data -> Select Cases -> Random sample of cases -> Sample -> Exactly -> 865 ... 885.

A amostra em estudo apresenta 8 variáveis que podem ser divididas em dois grupos:

Variáveis quantitativas:

CONS: consumo mensal de eletricidade (u.m.);

NUMP: número habitual de pessoas em casa;

AREA: área da casa;

REND: rendimento médio mensal da família (u.m.);

NUMC: número de crianças.

Variáveis qualitativas:

AC = 1 se a casa tem ar condicionado, 0 caso contrário;

APART = 1 se é apartamento, 0 caso contrário;

URB =1 se a casa está na zona urbana, 0 caso contrário.

Modelo

Modelo de Regressão Múltipla

Através do seguinte caminho: > Selecionar Analyze > Regression > Linear > selecionar ***CONS*** (Dependent), ***NUMP***, ***AREA***, ***REND***, ***NUMC***, ***AC***, ***APART***, ***URB*** (Block 1 of 1) > em Statistics selecionar Confidence intervals > Continue > OK, foi-nos possível estimar um modelo de regressão múltipla, ou seja, um modelo em que existe mais do que uma variável explicativa/independente, para podermos avaliar alguns determinantes do consumo de eletricidade.

O método de estimação usado em primeiro lugar foi o **ENTER,** que nos apresentou um modelo com todas as variáveis independentes como é possível observar na tabela 14:

Coeficientes[a]

Modelo		Coeficientes não padronizados		Coeficientes padronizados			95,0% Intervalo de Confiança para B		Estatísticas de colinearidade	
		B	Erro Erro	Beta	t	Sig.	Limite inferior	Limite superior	Tolerância	VIF
1	(Constante)	161,146	16,059		10,034	**<,001**	129,626	192,666		
	NUMP	5,245	1,164	,058	4,504	**<,001**	2,960	7,531	,998	**1,002**
	AREA	,051	,016	,042	3,211	**,001**	,020	,083	,994	**1,006**
	AC	166,366	4,868	,442	34,174	**<,001**	156,811	175,922	,991	**1,009**
	APART	56,426	4,685	,155	12,044	**<,001**	47,231	65,622	,999	**1,001**
	REND	,001	,000	,060	4,655	**<,001**	,001	,002	,995	**1,005**
	NUMC	91,154	2,510	,470	36,322	**<,001**	86,229	96,080	,993	**1,007**
	URB	255,510	4,816	,686	53,055	**<,001**	246,058	264,963	,993	**1,007**

a. Variável Dependente: CONS

Tabela 14 - Coeficientes pelo modelo ENTER

$$\boldsymbol{ONS} = {}_1 + {}_2\boldsymbol{NUMP} + {}_3\boldsymbol{AREA} + {}_4\boldsymbol{REND} + {}_5\boldsymbol{NUMC} + {}_6\boldsymbol{AC} + {}_7\boldsymbol{APART} + {}_8\boldsymbol{URB} +$$

$$\boldsymbol{ONS} = 161{,}146 + 5{,}245*\boldsymbol{NUMP} + 0{,}051*\boldsymbol{AREA} + 0{,}001*\boldsymbol{REND} + {}_5\boldsymbol{NUMC} + 166{,}366*\boldsymbol{AC} + 56{,}426*\boldsymbol{APART} + 255{,}510*\boldsymbol{URB} +$$

No entanto, como critério de confirmação para saber se o método **ENTER** nos apresentava o melhor modelo estimado, recorremos ao método **STEPWISE** de onde concluímos que incluir todas as variáveis seria a melhor opção, como se pode observar na tabela seguinte.

Coeficientes[a]

Modelo		Coeficientes não padronizados		Coeficientes padronizados			95,0% Intervalo de Confiança para B		Estatísticas de colinearidade	
		B	Erro Erro	Beta	t	Sig.	Limite inferior	Limite superior	Tolerância	VIF
1	(Constante)	448,861	7,493		59,907	<,001	434,155	463,567		
	URB	243,534	9,581	,654	25,418	<,001	224,729	262,339	1,000	1,000
2	(Constante)	354,716	7,449		47,620	<,001	340,095	369,336		
	URB	245,355	7,723	,659	31,768	<,001	230,196	260,514	1,000	1,000
	NUMC	86,903	4,025	,448	21,592	<,001	79,004	94,803	1,000	1,000
3	(Constante)	282,183	5,633		50,098	<,001	271,128	293,239		
	URB	256,071	5,324	,688	48,100	<,001	245,622	266,520	,996	1,004
	NUMC	91,043	2,772	,469	32,849	<,001	85,604	96,483	,998	1,002
	AC	166,889	5,384	,444	30,999	<,001	156,322	177,456	,994	1,006
4	(Constante)	255,395	5,709		44,733	<,001	244,189	266,601		
	URB	254,494	4,948	,684	51,438	<,001	244,783	264,205	,995	1,005
	NUMC	91,650	2,575	,473	35,587	<,001	86,595	96,705	,997	1,003
	AC	167,032	5,002	,444	33,396	<,001	157,216	176,849	,994	1,007
	APART	56,518	4,818	,156	11,731	<,001	47,062	65,974	,999	1,001
5	(Constante)	226,525	8,405		26,950	<,001	210,028	243,022		
	URB	255,423	4,894	,686	52,193	<,001	245,818	265,028	,993	1,007
	NUMC	91,377	2,546	,471	35,892	<,001	86,381	96,374	,997	1,003
	AC	167,412	4,944	,445	33,864	<,001	157,709	177,115	,993	1,007
	APART	56,541	4,762	,156	11,875	<,001	47,196	65,887	,999	1,001
	REND	,001	,000	,061	4,634	<,001	,001	,002	,998	1,002
6	(Constante)	201,596	10,016		20,127	<,001	181,936	221,255		
	URB	255,759	4,841	,687	52,827	<,001	246,257	265,261	,993	1,007
	NUMC	91,624	2,519	,472	36,374	<,001	86,680	96,568	,996	1,004
	AC	166,944	4,891	,444	34,131	<,001	157,344	176,544	,993	1,007
	APART	56,616	4,710	,156	12,021	<,001	47,372	65,861	,999	1,001
	REND	,001	,000	,059	4,534	<,001	,001	,002	,996	1,004
	NUMP	5,225	1,171	,058	4,463	<,001	2,927	7,523	,998	1,002
7	**(Constante)**	**161,146**	**16,059**		**10,034**	**<,001**	**129,626**	**192,666**		
	URB	**255,510**	**4,816**	**,686**	**53,055**	**<,001**	**246,058**	**264,963**	**,993**	**1,007**
	NUMC	**91,154**	**2,510**	**,470**	**36,322**	**<,001**	**86,229**	**96,080**	**,993**	**1,007**
	AC	**166,366**	**4,868**	**,442**	**34,174**	**<,001**	**156,811**	**175,922**	**,991**	**1,009**
	APART	**56,426**	**4,685**	**,155**	**12,044**	**<,001**	**47,231**	**65,622**	**,999**	**1,001**
	REND	**,001**	**,000**	**,060**	**4,655**	**<,001**	**,001**	**,002**	**,995**	**1,005**
	NUMP	**5,245**	**1,164**	**,058**	**4,504**	**<,001**	**2,960**	**7,531**	**,998**	**1,002**
	AREA	**,051**	**,016**	**,042**	**3,211**	**,001**	**,020**	**,083**	**,994**	**1,006**

a. Variável Dependente: CONS

Tabela 15 - Modelos estimados pelo método STEPWISE

Ao analisar os modelos estimados pelo método STEPWISE, destacamos que o melhor modelo é o último da tabela, modelo este que utiliza todas as variáveis. Neste caso coincide com o método ENTER, o que nos levou a optar por utilizar este.

Resumo do modelo

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa
1	,926[a]	,858	,856	68,785287698

a. Preditores: (Constante), URB, NUMC, NUMP, APART, REND, AREA, AC

Tabela 16 - Resumo do modelo

Como o coeficiente de determinação ajustado (R) possui um valor próximo de 1, podemos constatar que as variáveis independentes ajudam a explicar a variação da variável dependente.

As variáveis inseridas explicam, em conjunto, cerca de **85,8%** da variação do consumo de eletricidade.

Inferência estatística

Através do teste ANOVA podemos testar a significância global do modelo, ou seja, verificar se pelo menos uma das variáveis independentes tem um efeito significativo sobre a variável dependente (***CONS***). Supomos assim as seguintes hipóteses:

H_0: $_2 = _3 = _4 = _5 = _6 = _7 = _8 = 0$ (modelo globalmente não adequado)

H_1: $_j \neq 0$ (modelo globalmente adequado)

ANOVA[a]

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	24410817,753	7	3487259,679	737,044	<,001[b]
	Resíduo	4054823,344	857	4731,416		
	Total	28465641,097	864			

a. Variável Dependente: CONS

b. Preditores: (Constante), URB, NUMC, NUMP, APART, REND, AREA, AC

Tabela 17 - ANOVA

Como podemos ver na tabela 17, temos um $valor\text{-}p = sig \approx 0,0 < \alpha = 0,05$, portanto **rejeitamos H_0** e concluímos que **pelo menos uma das variáveis independentes afeta o consumo de electricidade**. Assim sendo, vamos averiguar a significância individual das variáveis para ver, qual ou quais, é que podem afetar a variável dependente. Testamos, para cada uma das variáveis independentes, as seguintes hipóteses:

H_0: $_j = 0$ (a variável é estatisticamente não relevante), $j \quad [2, 8]$

H_1: $_j \neq 0$ (a variável é estatisticamente relevante)

Após analisar a tabela 14, constatamos que o $valor\text{-}p = sig \approx 0,0 < \alpha = 0,05$ **em todas as variáveis** e por isso, **rejeitamos H_0**, ou seja, **todas as variáveis consideradas neste modelo afetam significativamente o consumo de eletricidade**. Podemos ainda, através dos intervalos de confiança, chegar à mesma conclusão. Ou seja, como 0 não pertence a nenhum dos intervalos evidenciados constata-se que $_j$ não pode ser igual a 0 e por isso rejeita-se a hipótese nula.

Deste modo, podemos dizer que, supondo iguais condições para as restantes variáveis, estimamos que:

- O aumento de uma unidade no número de pessoas habitualmente em casa (***NUMP***), implica um aumento médio estimado de 5,245 u.m. no consumo de eletricidade;
- O aumento de uma unidade adicional na área da casa (***AREA***), implica um aumento médio estimado de 0,051 u.m. no consumo de eletricidade;

- O aumento de uma unidade adicional no rendimento médio mensal da família (***REND***), implica um aumento médio estimado de 0,001 u.m. no consumo de eletricidade;
- O aumento de uma unidade adicional no número de crianças em casa (***NUMC***), implica um aumento médio estimado de 91,154 u.m. no consumo de eletricidade;
- Se a casa tiver ar condicionado (***AC*** = 1), o consumo de eletricidade vai ser maior em 166,366 u.m. comparado com o consumo de uma casa sem ar condicionado (***AC*** = 0)
- Se a casa for um apartamento (***APART*** = 1), o consumo de eletricidade vai ser maior em 56,426 u.m. comparado com o consumo de uma casa que não o seja (***APART*** = 0)
- Se a casa estiver situada numa zona urbana (***URB*** = 1), o consumo de eletricidade vai ser maior em 255,510 u.m. comparado com o consumo de uma casa que não está localizada nesse tipo de zona (***URB*** = 0).

Pressupostos do Modelo

Para se garantir a validade e a confiabilidade dos resultados obtidos com o modelo anteriormente estimado temos de garantir alguns pressupostos. Como se trata de um modelo de regressão linear múltipla devemos ter em consideração:

Independência dos erros;
Variância dos erros constante (homocedasticidade);
Normalidade da distribuição dos erros;
Ausência de multicolinearidade.

Assim sendo, para assegurar a validade dos pressupostos vamos efetuar uma análise aos resíduos do modelo uma vez que se tratam das estimativas dos erros ( ). O termo erro é uma variável não observável e que considera fatores que podem afetar o comportamento da variável explicada e que não são considerados no modelo.

A nossa verificação vai então ser feita com base nos outputs obtidos pelo caminho: Analyze > Regression > Linear > Em Plots escolher ZRESID (Y) e ZPRED (X) > Continue > OK.

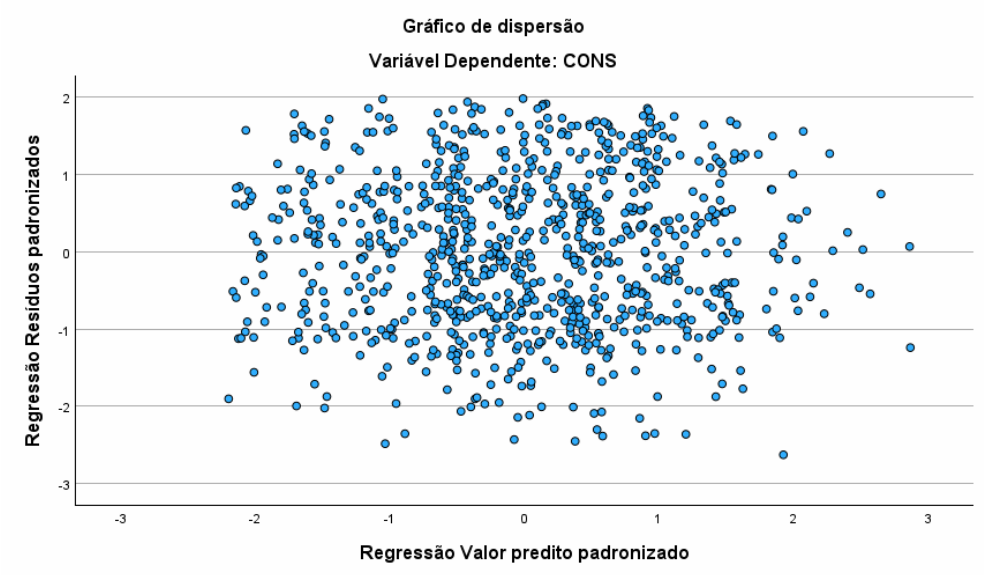


Gráfico 6- Dispersão dos Resíduos

Através do gráfico 6 é possível observar que existe uma **mancha de pontos aleatórios e sem nenhum tinho de padrão** o que nos leva a **validar a hipótese de independência dos erros**. No que concerne ao pressuposto da variância dos erros, pode também este ser comprovado uma vez que é visível uma **dispersão de pontos praticamente igual ao longo do eixo dos xx**. **Valida-se** por isso a **hipótese de homocedasticidade**.

Resumo de processamento de casos

	Casos					
	Válido		Omisso		Total	
	N	Porcentagem	N	Porcentagem	N	Porcentagem
Unstandardized Residual	865	100,0%	0	0,0%	865	100,0%

Tabela 18- Resíduos

Testes de Normalidade

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	gl	Sig.	Estatística	gl	Sig.
Unstandardized Residual	,053	865	**<,001**	,984	865	<,001

a. Correlação de Significância de Lilliefors

Tabela 19- Teste de Normalidade dos Resíduos

Em relação ao pressuposto da normalidade dos erros, através do teste Kolmogorov-Smirnov, testamos as seguintes hipóteses:

$\boldsymbol{H_0}$: Os erros seguem uma distribuição normal

$\boldsymbol{H_1}$: Os erros não seguem uma distribuição normal

Concluímos que o ***valor-p = sig ≈ 0,0 < alfa = 0,05*** (tabela 19) e por isso se rejeita a hipótese nula (H_0), ou seja, os erros não seguem uma distribuição normal.

Para analisarmos a multicolinearidade são observadas as medidas **tolerância** e **VIF**. Tendo em consideração que, a tolerância é a percentagem da variação da variável independente que não é explicada pelas restantes variáveis independentes e que VIF é o inverso disso podemos concluir que, para valores de tolerância e VIF mais próximos de 1 menor a multicolinearidade. Observa-se assim na tabela 14, que os **valores para tais medidas são próximos de 1**, o que nos leva a concluir que **não existe multicolinearidade**.

Podemos concluir que todos os pressupostos se cumprem com a exceção da normalidade da distribuição dos resíduos, o que pode ter implicação para prever e estimar os coeficientes. No entanto, como a nossa amostra tem uma dimensão grande, a falta de normalidade tem pouco impacto prático na análise e o modelo ainda pode ser considerado confiável na maioria dos propósitos.

Conclusão

Em síntese, este relatório estatístico apresenta uma análise detalhada de duas bases de dados: a satisfação dos passageiros de uma companhia aérea e os fatores que influenciam o consumo de eletricidade em residências. Utilizando métodos rigorosos de análise de dados, a primeira parte do relatório aplicou os testes de Mann-Whitney U e Kruskal-Wallis para identificar diferenças significativas na percepção de conforto e satisfação dos serviços em diferentes tipos de viagem e classes. A análise fatorial permitiu reduzir as variáveis de satisfação a componentes principais que, em conjunto, explicam 69% da variância total, destacando fatores essenciais como conforto a bordo e conveniência de serviços.

Na segunda parte, a análise de regressão múltipla identificou os principais determinantes do consumo de eletricidade, validando o modelo de forma robusta. O tamanho considerável da amostra confere validade aos resultados, que evidenciam a influência de variáveis como o número de moradores, ar condicionado e localização urbana.

Assim, foram apresentadas conclusões fulcrais para o planeamento de melhorias nos serviços de transporte aéreo e para o entendimento dos fatores que impactam o consumo de eletricidade, com potencial para orientar decisões estratégicas, assim como facilitar a sua interpretação.