

UFSCar - UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA

Trabalho de Planejamento

O Efeito da Prática de Alguns Esportes na Gordura Corporal

Gabriela Pereira Soares RA: 758549

Henrique Salviano RA: 614440

Luis Roberto Ferreira Junior RA: 744864

Rhayani Aparecida Paiuta RA: 758584

São Carlos - SP

Dezembro de 2020

Conteúdo

1	Introdução	2
1.1	Apresentação do problema	2
1.2	Objetivo	2
2	Conjunto de Dados	3
3	Metodologia	4
3.1	Modelo	4
3.2	Tabela de ANOVA	4
4	Análise Descritiva	6
5	Análise de Diagnóstico	9
5.1	Verificação da existência de interação entre bloco e fator	9
5.2	Verificação da independência dos erros	10
5.3	Verificação da normalidade dos erros	12
5.4	Verificação da homocedasticidade dos erros	13
6	Teste de hipótese para o efeito principal do fator Esporte	14
7	Comparações Múltiplas	16
8	Conclusão	17
	Referências	18

1 Introdução

1.1 Apresentação do problema

A prática de atividade físicas e o treino de esportes são essenciais para a saúde e podem reduzir a gordura corporal. Algumas questões como o tipo de atividade praticada, frequência e intensidade do exercício podem proporcionar maior ou menor queima de gordura do corpo.

Muitas pesquisas têm sido realizadas com o interesse em identificar quais modalidades esportivas apresentam melhores resultados na perda de peso, visto que a procura por uma qualidade de vida mais saudável e um corpo mais definido é um assunto recorrente na sociedade contemporânea.

Deste modo, uma academia ou competição esportiva podem estar interessadas em traçar um perfil de seus atletas ou competidores em relação à porcentagem de gordura corporal, e identificar quais das modalidades esportivas oferecidas apresentam indivíduos com maior gordura média, levando em consideração as características físicas e biológicas de cada um.

1.2 Objetivo

Dado o interesse de academias ou competições conhecerem a gordura corporal de seus atletas, o objetivo desse trabalho é analisar se existe impacto da prática de alguns esportes (corrida de barco, natação e tênis) na gordura corporal média. Além disso, caso o impacto do esporte seja significativo, será analisado qual modalidade impacta de forma mais significativa na gordura corporal.

2 Conjunto de Dados

O conjuntos de dados desse trabalho foi utilizado em um trabalho da matéria Análise de Regressão de um integrante do grupo. O conjunto de dados contém informações a respeito de 77 indivíduos e contém 13 variáveis com informações biológicas destes indivíduos, sendo duas delas qualitativas.

Com o objetivo de construir um experimento em blocos completamente aleatorizado com efeito fixo e balanceado, o grupo decidiu trabalhar com as duas variáveis qualitativas presentes no conjuntos de dados, a variável Gênero que tem dois níveis (Masculino e Feminino) e a variável Esporte, que inicialmente apresentava 10 modalidade diferentes, mas que para fins do experimento fixamos três níveis (Corrida de Barco, Natação e Tênis).

Como sabemos, homens e mulheres apresentam diferenças biológicas entre si, uma delas é a gordura corporal. Juntando essa informação com a tarefa de construir um experimento em blocos, o grupo escolheu como variável resposta a Gordura Corporal, medida em quilogramas, e, conseqüentemente, definiu como bloco o Gênero, pois essa variável seria uma restrição natural e uma fonte de variação para a gordura corporal. Dessa forma, a variável Esporte será entendida como Fator e o objetivo principal do trabalho será analisar se existe efeito do Esporte na gordura corporal dos indivíduos.

Posteriormente, para que o Experimento fosse completamente aleatorizado e balanceado, sorteamos uma amostra de tamanho quatro ($n = 4$) para cada combinação de Gênero e Esporte (total de 6 combinações). É importante ressaltar que o grupo escolheu o número de réplicas $n = 4$ pois na combinação (Masculino, Tênis) o conjunto de dados apresentava apenas 4 observações. Logo, como o experimento deveria ser balanceado, todos as combinações receberam um número de réplicas igual a 4. Os dados que utilizamos na análise estão apresentados na Tabela 1.

Tabela 1: Conjunto de dados amostrado.

Esporte Gênero	Barco		Natação		Tênis	
Feminino	78.7	79.7	59.0	69.7	56.0	61.1
	71.3	80.5	63.9	60.0	69.7	47.4
Masculino	89.5	92.3	83.0	74.4	80.0	73.8
	85.4	87.9	78.0	78.0	71.1	76.7

3 Metodologia

3.1 Modelo

Como iremos realizar um experimento em blocos completamente aleatorizado com efeito fixo e balanceado, será utilizado o modelo a seguir:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}. \quad (1)$$

Em que,

- Y_{ijk} é a variável resposta, $i = 1, 2, 3$, $j = 1, 2$, $k = 1, 2, 3, 4$;
- μ é o efeito comum que independe de blocos ou tratamentos;
- τ_i é o efeito do i -ésimo tratamento (nível do fator A), $i = 1, 2, 3$;
- β_j é o efeito do j -ésimo nível do bloco, $j = 1, 2$;
- ε_{ijk} são os erros aleatórios do modelo, $i = 1, 2, 3$, $j = 1, 2$, $k = 1, 2, 3, 4$.

Suposições: $\varepsilon_{ijk} \sim N(0, \sigma^2)$ e são independentes; $(\tau\beta)_{ij} = 0$.

Restrições:

$$\sum_{i=1}^a \tau_i = 0, \quad \sum_{j=1}^b \beta_j = 0. \quad (2)$$

Parâmetros e Estimadores:

Tabela 2: Parâmetros do modelo e seus respectivos estimadores.

Parâmetro	Estimador
μ	$\hat{\mu} = \bar{y}_{...}$
τ_i	$\hat{\tau}_i = \bar{y}_{i..} - \bar{y}_{...}$
β_j	$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$

3.2 Tabela de ANOVA

Para analisarmos se existe efeito da prática de um determinado esporte na gordura corporal de um indivíduo, iremos conduzir testes de hipóteses através da Análise de Variância (ANOVA). Para este experimento, a tabela ANOVA é dada por:

Tabela 3: Tabela ANOVA para o modelo utilizado.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F
B (Blocos)	$b - 1$	SQB	QMB	$\frac{QMB}{QME}$
A (Tratamentos)	$a - 1$	SQA	QMA	$\frac{QMA}{QME}$
Erro	$nab - (a + b - 1)$	SQE	QME	
Total	$N - 1$	SQT		

Em que,

- $a = 3$ é o número de níveis do Fator A (tratamentos);
- $b = 2$ é o número de níveis do bloco;
- $n = 4$ é o número de unidades experimentais por tratamento em cada bloco;
- $N = nab = 24$: número total de unidades experimentais no experimento;
- $SQT = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{...})^2$ é a soma de quadrados total;
- $SQA = bn \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2$ é a soma de quadrados do Fator A;
- $SQB = an \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2$ é a soma de quadrados do bloco B;
- $SQE = SQT - SQA - SQB$ é a soma de quadrados dos erros;
- $QMA = \frac{SQA}{a-1}$ é o quadrado médio do Fator A;
- $QMB = \frac{SQB}{b-1}$ é o quadrado médio do Fator B;
- $QME = \frac{SQE}{nab-(a+b-1)}$ é o quadrado médio dos erros.

4 Análise Descritiva

Antes da aplicação dos métodos inferenciais, será realizada uma breve análise descritiva dos dados a fim de explorar as suposições que eventualmente foram feitas na definição do modelo e, além disso, apresentar ao leitor uma breve noção dos resultados que os testes de hipótese podem trazer. Lembrando que durante essa análise nenhuma conclusão será tomada, as conclusões finais serão tomadas a partir dos métodos inferenciais utilizados na Seção 6.

Como dito anteriormente, um experimento em blocos completamente aleatorizado com efeito fixo e balanceado será conduzido para esse conjunto de dados, sendo a variável resposta (Y) a gordura corporal, o bloco como o gênero da unidade experimental (Feminino e Masculino) e o fator como o esporte praticado pela unidade experimental (Corrida de Barco, Natação e Tenis).

Sabemos que o bloco deve ser representado por uma característica que divide “naturalmente” os indivíduos, ou seja, o bloco funciona como uma restrição na qual dentro do mesmo bloco os indivíduos são homogêneos e entre cada bloco os indivíduos são heterogêneos. Nas figuras 1a e 1b, dois gráficos são apresentados:

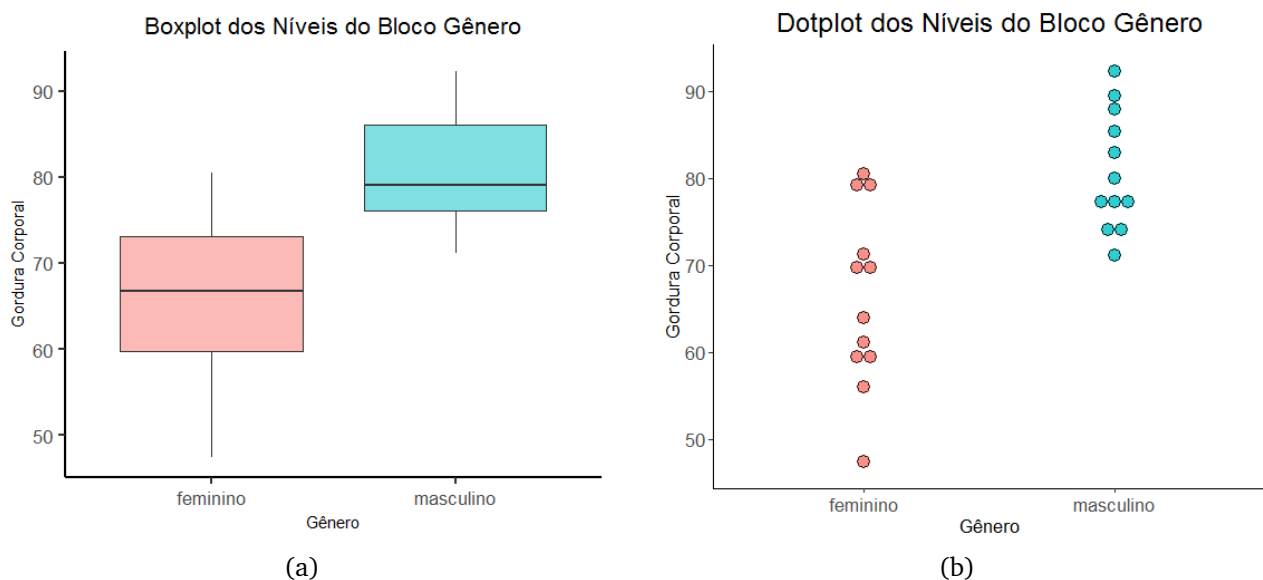


Figura 1: Boxplot e Dotplot para o bloco Gênero.

Observe que, tanto pelos boxplots quanto pelos dotplots, as unidades experimentais do sexo masculino tem gordura corporal relativamente maior que as unidades experimentais do sexo feminino. Ou seja, os gráficos explicitam o significado de bloco, para ser possível

conduzir esse experimento é de fato necessário que os dois blocos apresentem diferença entre si com relação a variável resposta gordura corporal, e como sabemos essa diferença entre blocos se dá pela forma biológica.

Agora, vamos analisar o comportamento da variável gordura corporal com relação aos esportes praticados pelas unidades experimentais. Como dito anteriormente, o esporte nesse modelo é entendido como um fator (fixo), ou seja, queremos analisar o efeito da prática de três esportes na gordura corporal. As Figuras 2a e 2b trazem o boxplot e dotplot, respectivamente, para essa situação:

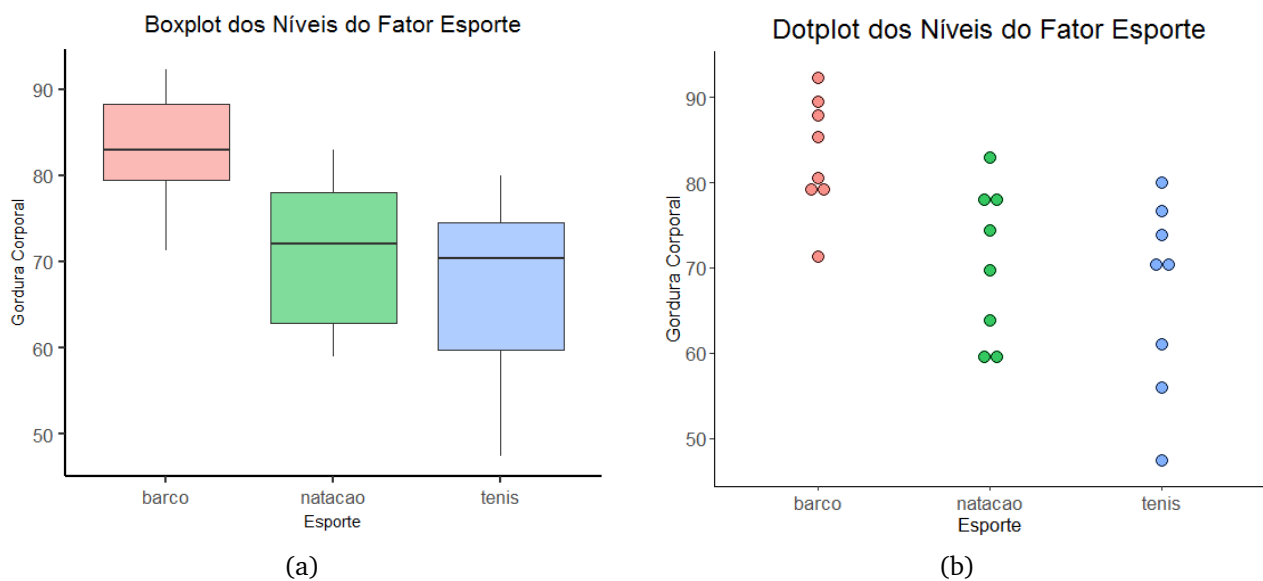


Figura 2: Boxplot e Dotplot para o fator Esporte.

Visivelmente podemos notar que a mediana da gordura corporal dos indivíduos que praticam corrida de barco é maior do que os indivíduos que praticam natação ou tênis. Existe uma grande intersecção entre as caixas dos boxplots dos indivíduos que praticam natação e tênis, o que pode indicar que a gordura corporal das indivíduos que praticam esses dois esportes não diferencie muito entre si. A ideia de que a gordura corporal dos indivíduos que praticam corrida de barco seja maior também pode ser analisada pelo dotplot.

E por fim, como tratamos a variável gênero como bloco, não faz sentido acrescentar nesse modelo o efeito de interação entre gênero e esporte, uma vez que naturalmente existe uma diferença entre homens e mulheres com relação a gordura corporal. Dessa forma, a Figura 3 traz o gráfico de interação para nosso conjunto de dados:

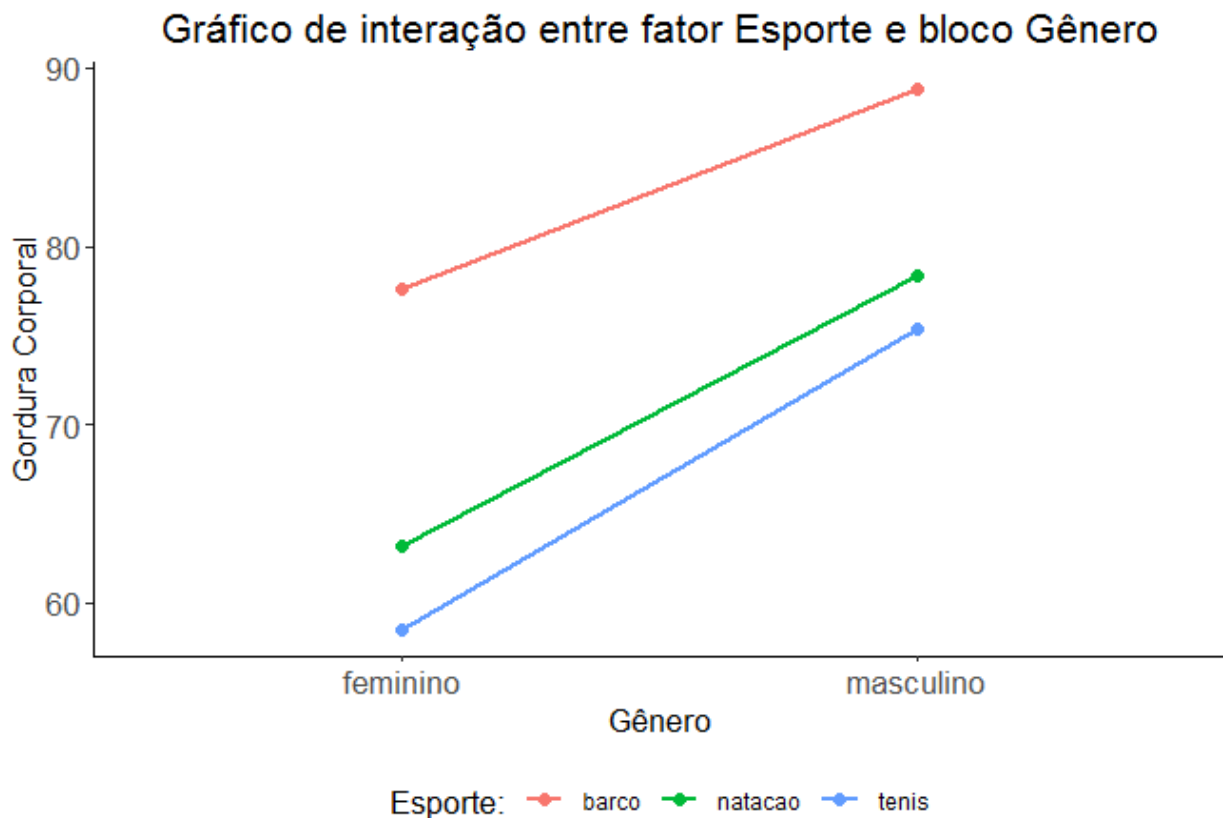


Figura 3: Gráfico de interação entre fator Esporte e bloco Gênero.

O gráfico esboça três linhas quase paralelas, sugerindo que não existe efeito de interação entre o gênero do indivíduo e o esporte que o mesmo pratica. Além disso, o gráfico também sugere que a gordura corporal dos indivíduos que praticam corrida de barco parece ser maior do que a gordura corporal dos indivíduos que praticam os outros esportes (a reta rosa está distante das retas verde e azul).

Ainda sem realizar a análise de diagnóstico (verificar as suposições do modelo sobre os erros), a análise descritiva já nos sugere que o experimento em blocos é adequado para esses dados, uma vez que o bloco gênero cumpre com seu papel de separar naturalmente indivíduos que apresentam diferenças biológicas e, além disso, gênero e esporte não parecem ter efeito de interação entre si, como é recomendado para a construção de um modelo desse tipo.

5 Análise de Diagnóstico

Antes de iniciarmos a análise metodológica baseada em testes de hipóteses, devemos averiguar primeiro se o modelo descrito na Equação 1 é adequado para este caso.

Para verificar a adequabilidade do modelo baseado em um delineamento em blocos completamente aleatorizado, devemos verificar se as suposições de independência, normalidade e homocedasticidade dos erros são satisfeitas.

Para isto, faremos algumas análises gráficas e testes de hipóteses baseados nos resíduos do modelo, $\hat{\epsilon}_{ijk} = r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. Se as condições de normalidade, independência e homocedasticidade dos mesmos forem satisfeitas, então estenderemos estas mesmas conclusões para os erros aleatórios.

Além disso, devemos verificar se de fato não existe efeito de interação entre o bloco Gênero e o fator Esporte.

5.1 Verificação da existência de interação entre bloco e fator

Vimos na análise descritiva que parece de fato não haver efeito de interação entre o fator Esporte e o bloco Gênero, devido às retas estarem praticamente paralelas. Neste caso, podemos ajustar o modelo contendo o efeito de interação entre bloco e fator, adicionando a componente $(\tau\beta)_{ij}$ ao modelo descrito na Equação 1, e testar se de fato existe efeito de interação ou não.

Vale ressaltar que, para realizar o teste de hipóteses, fizemos previamente uma breve análise dos resíduos do modelo com interação e averiguamos que todas as suposições foram atendidas.

Portanto, teremos as seguintes hipóteses:

$$\begin{cases} H_0 : (\tau\beta)_{ij} = 0, \quad \forall i = 1, 2, 3 \quad j = 1, 2 \text{ (Não existe efeito de interação entre bloco e fator);} \\ H_1 : (\tau\beta)_{ij} \neq 0, \quad \text{para pelo menos um par } i, j \text{ (Existe efeito de interação entre bloco e fator).} \end{cases}$$

Fixaremos um nível de significância de $\alpha = 5\%$.

A estatística do teste e sua distribuição, sob H_0 , é dada por:

$$F_{AB} = \frac{QMAB}{QME} \sim F_{2,18}. \quad (3)$$

Rejeitaremos a hipótese nula se $F_{AB} > F_{0.95,2,18} = 3.554557$.

Tabela 4: Tabela ANOVA do modelo com interação.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F	P-valor
Gênero	1	1248,5	1248,5	45,635	$2,49 \cdot 10^{-6}$
Esporte	2	1147,6	573,8	20,974	$1,98 \cdot 10^{-5}$
Gên:Esp	2	33,4	16,7	0,611	0,554
Resíduos	18	492,4	27,4		
Total	23	2921,4			

Utilizando a tabela de ANOVA para o modelo com interação, obtemos

$$f_{AB} = \frac{16.7}{27.4} = 0.611. \quad (4)$$

Como $f_{AB} < 3.554557$, não rejeitamos H_0 . Ao nível de significância de 5%, não há evidências de que exista efeito de interação entre o fator Esporte e o bloco Gênero.

Além disso, vemos pela Tabela 4 que o p-valor relacionado ao efeito do bloco gênero é bem pequeno, nos dando indícios de que de fato, ao nível de significância, há diferença entre os gêneros em relação à gordura corporal, como já esperávamos na análise descritiva.

Considerando, então, a análise gráfica e o resultado do teste, seguiremos com a verificação das suposições dos erros utilizando o modelo sem interação descrito na Equação 1.

5.2 Verificação da independência dos erros

Como dito anteriormente, homens e mulheres de cada modalidade esportiva foram sorteados aleatoriamente do conjunto de dados completo. Portanto, não temos indícios de que haja dependência dos resíduos relacionados às observações.

Para consolidar esta conclusão, podemos construir um gráfico dos resíduos versus a ordem de coleta das observações:

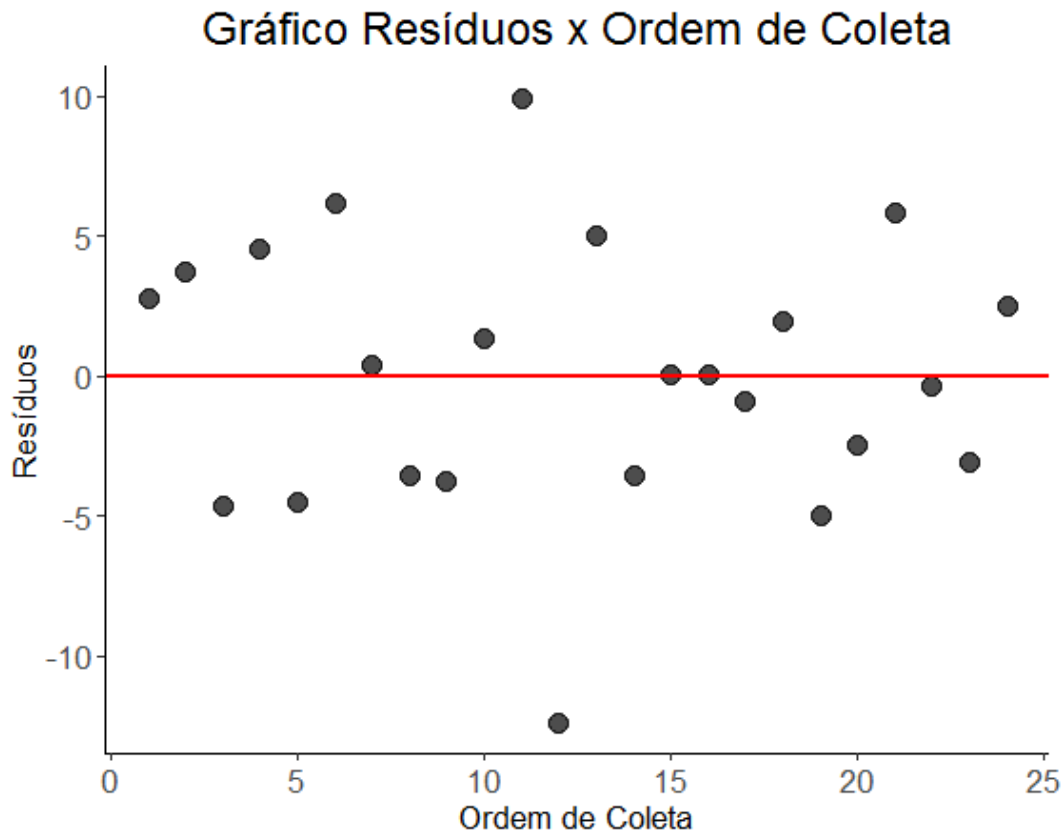


Figura 4: Gráfico de resíduos versus ordem de coleta.

Pela Figura 4, podemos notar que os pontos estão aleatoriamente distribuídos em torno da reta vermelha, onde $r_{ij} = 0$. Não notamos também nenhum tipo de comportamento padrão entre as observações em relação à ordem de coleta.

O teste de autocorrelação de Durbin Watson foi executado para uma verificação quantitativa dessa suposição, onde fixamos um nível de significância de $\alpha = 5\%$ e as hipóteses:

$$\begin{cases} H_0 : \rho = 0 & \text{(Não há presença de autocorrelação nos resíduos);} \\ H_1 : \rho \neq 0 & \text{(Há presença de autocorrelação nos resíduos).} \end{cases}$$

O p-valor obtido foi de 0.9873. Deste modo, como $0.9873 > 0.05$, não rejeitamos H_0 . Ao nível de significância de 5%, não há evidências de que exista autocorrelação entre os resíduos.

Portanto, não temos evidências de que os resíduos são dependentes.

5.3 Verificação da normalidade dos erros

Para verificar se a suposição de normalidade é satisfeita, podemos analisar o gráfico normal probabilístico dos resíduos.

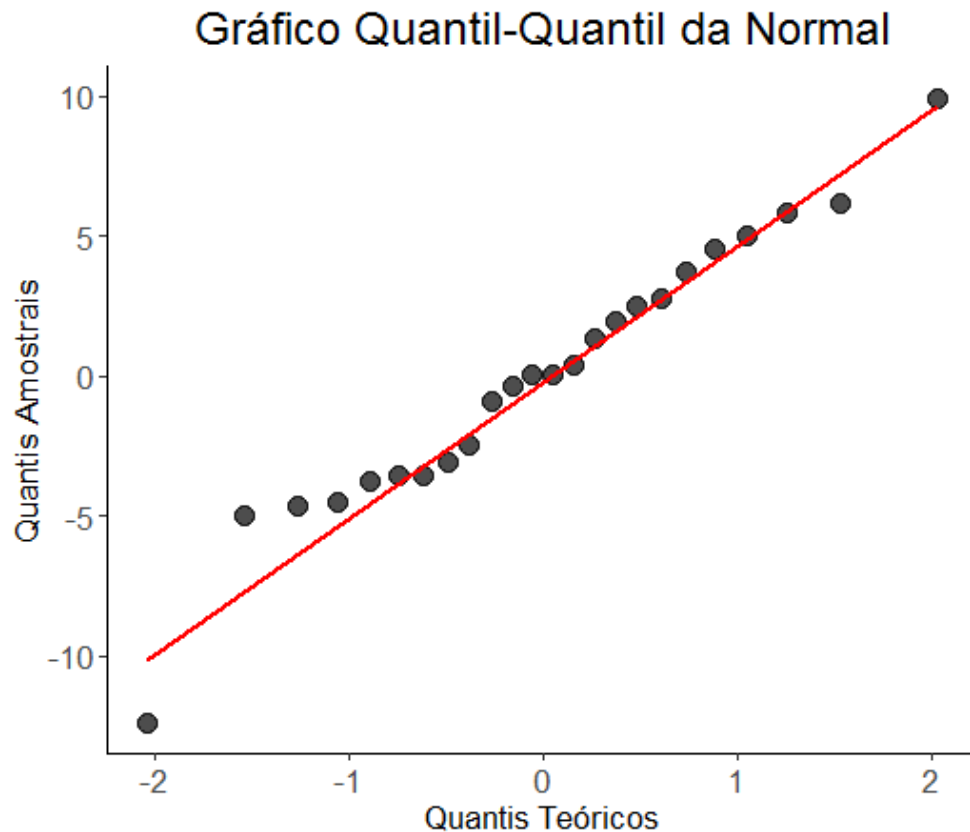


Figura 5: Gráfico normal probabilístico para os resíduos.

Pela Figura 5, percebemos que os pontos encontram-se bem próximos da linha vermelha. Entretanto, à esquerda notamos alguns pontos se deslocando um pouco mais da reta do que os demais pontos, mas aparentemente nada muito discrepante.

Para averiguar de fato a normalidade dos resíduos, podemos realizar uma série de testes de hipóteses e avaliar o p-valor correspondente de cada um.

Para todos os testes a seguir, consideraremos um nível de significância de $\alpha = 5\%$ e as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Os resíduos seguem distribuição normal;} \\ H_1 : \text{Os resíduos não seguem distribuição normal.} \end{cases}$$

Obtemos a tabela a seguir contendo o p-valor de cada teste:

Tabela 5: Resultados dos testes de normalidade.

Teste	Lilliefors	Kolmogorov Smirnov	Shapiro Wilk	Shapiro Francia	Anderson Darling
P-valor	0.6733	0.9447	0.7249	0.4159	0.7017

Como os p-valores > 0.05 para todos os testes utilizados na Tabela 5, não rejeitamos H_0 . Portanto, ao nível de significância de 5%, não há evidências de que os resíduos não seguem distribuição normal.

5.4 Verificação da homocedasticidade dos erros

Por fim, para verificar se a suposição de homocedasticidade dos erros é válida, podemos analisar o gráfico de resíduos contra valores preditos, dado abaixo:

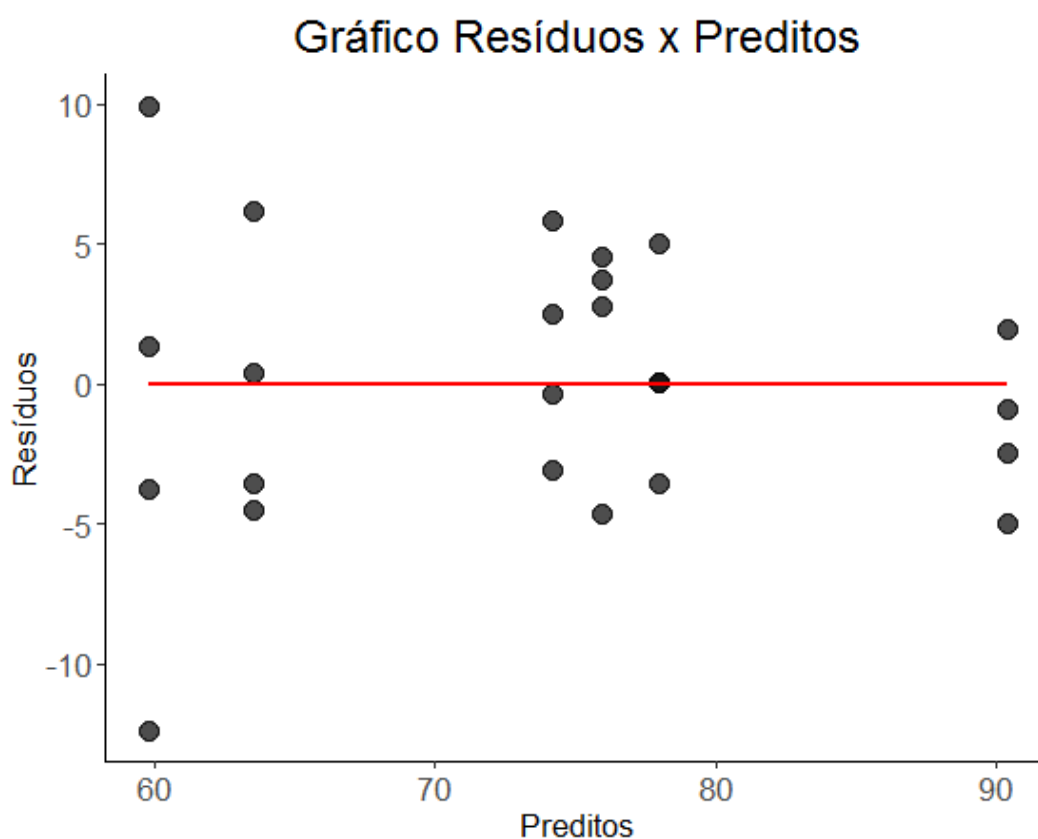


Figura 6: Gráfico resíduos versus valores preditos.

Pela Figura 6, podemos notar que os pontos encontram-se distribuídos de forma homogênea em torno da reta vermelha em relação aos grupos, com exceção da primeira coluna de pontos, onde parecem estar mais espaçados do que os demais.

Podemos realizar, também, testes de hipóteses para verificar a homogeneidade das variâncias dos resíduos.

Fixaremos, portanto, um nível de significância de $\alpha = 5\%$, e testaremos as seguintes hipóteses:

$$\begin{cases} H_0 : \sigma_{ij}^2 = \sigma^2, \quad \forall i = 1, 2, 3 \quad j = 1, 2 \text{ (Os resíduos possuem variância constante);} \\ H_1 : \sigma_{ij}^2 \neq \sigma^2, \quad \text{para ao menos um par } i, j \text{ (Os resíduos não possuem variância constante).} \end{cases}$$

Os resultados obtidos para cada teste e o respectivos p-valores encontram-se na Tabela 6:

Tabela 6: Resultados dos testes de homocedasticidade.

Teste	Bartlett	Hartley	Levene (média)	Levene (mediana)
P-valor	0.3762	0.4662	0.221	0.2885

Como p-valor > 0.05 para todos os testes utilizados na Tabela 6, não rejeitamos H_0 . Ao nível de significância de 5%, não há evidências de que os resíduos sejam heterocedásticos.

Portanto, temos que todas as suposições para o modelo são satisfeitas, e podemos seguir com a análise metodológica.

6 Teste de hipótese para o efeito principal do fator Esporte

Inicialmente, conduziremos um teste de hipóteses associado ao efeito principal do fator A para verificar se há diferença entre os 3 tipos de esporte em reação à gordura média corporal dos indivíduos.

Hipóteses:

$$\begin{cases} H_0 : \tau_i = 0 \text{ para todo } i = 1, 2, 3 \text{ (A gordura média corporal é a mesma para os 3 esportes);} \\ H_1 : \tau_i \neq 0 \text{ para pelo menos um } i \text{ (A gordura média corporal não é a mesma para os 3 esportes).} \end{cases}$$

Nível de significância:

Utilizaremos um nível de significância de $\alpha = 5\%$.

Estatística do teste e sua distribuição sob H_0 :

Como o fator Esporte tem 3 níveis ($a = 3$), e temos 2 blocos ($b = 2$), e quatro unidades experimentais por esporte em cada gênero ($n = 4$), então

$$(a - 1) = 2 \text{ e } nab - (a + b - 1) = 20.$$

Assim, sob H_0 ,

$$F_A = \frac{QMA}{QME} \sim F_{2,20}. \quad (5)$$

Região crítica:

Considerando que, nesse teste, rejeita-se H_0 apenas para valores grandes de F_0 , obtém-se utilizando o software RStudio para $\alpha = 5\%$,

$$RC : F_A > 3.49. \quad (6)$$

Tabela ANOVA:

Tabela 7: Tabela ANOVA para o modelo estimado.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F	Valor-p
Gênero	1	1248.5	1248.5	47.48	0.00000108
Esporte	2	1147.6	573.8	21.82	0.00000939
Erro	20	525.9	26.3		
Total	23	2922			

Como $f_A = 21.82$ pertence à região crítica, rejeitamos H_0 . Utilizando o valor-p, também rejeitamos H_0 , pois o *valor - p* = 0.00000939 é menor do que o nível de significância de $\alpha = 5\%$. Assim ao nível de significância de $\alpha = 5\%$, há evidências de que a gordura média corporal não é a mesma para todos os esportes.

Para complementar a análise, como concluímos que a gordura média corporal não é

a mesma para todos os esportes, necessitamos realizar comparações múltiplas para estudar quais esportes, em média, diferem.

7 Comparações Múltiplas

Utilizaremos o método de Tukey para calcular os intervalos de $100(1 - \alpha)\%$ de confiança para as diferenças de médias, haja vista que esse método produz intervalos mais estreitos e, também, por que queremos realizar todas as comparações múltiplas dois a dois. Os intervalos de confiança são dados pela Equação 7.

$$IC[\mu_i - \mu_l; 100(1 - \alpha)\%] = \left[(\bar{Y}_i - \bar{Y}_l) \pm q_\alpha(a, N - a) \sqrt{\frac{QME}{n}} \right]. \quad (7)$$

em que $q_\alpha(a, N - a)$ é o valor da Tabela da amplitude total studentizada (Tabela de Tukey), referente aos a número de tratamentos, $(N - a)$ graus de liberdade associados ao erro e com nível de significância α .

Utilizando o software *R Studio*, obtivemos os intervalos de 95% confiança para a diferença entre as médias de cada tratamento. Os intervalos estão dispostos na Tabela 8.

Tabela 8: Intervalos de confiança de Tukey.

Comparação	Diferença	Limite inferior	Limite superior	p-valor
Natação-Barco	-12.4125	-18.89914	-5.925859	0.0002802
Tênis-Barco	-16.1875	-22.67414	-9.700859	0.0000106
Tênis-Natação	-3.7750	-10.26164	2.711641	0.3248704

Pela Tabela 8, concluímos que, com 95% de confiança, a gordura corporal dos atletas de corrida de barco é, em média, maior que a dos atletas de natação e de tênis, confirmando, assim, o que havíamos previsto na Seção 4. Também podemos concluir que, com 95% de confiança, não há diferença, em média, na gordura corporal dos atletas de tênis e de natação, assim como também havíamos previsto na Seção 4.

8 Conclusão

Diante de tudo que foi exposto, constatamos, primeiramente, na Seção 4, que o bloco Gênero realmente é uma fonte de variação que torna as unidades experimentais heterogêneas entre si em relação à gordura corporal dos atletas, ou seja, as mulheres da amostra, em média, possuem gordura corporal menor em relação aos homens, assim como havíamos sugerido, por motivos biologicamente explicados, nas Seções 1.1 e 1.2. Portanto, nos baseando na análise descritiva e também no p-valor obtido na Tabela de ANOVA da Seção 5.1, o delineamento em blocos completamente aleatorizado, utilizando a variável Gênero como bloco, é adequado para testar o efeito principal do fator Esporte na gordura corporal dos atletas.

Além disso, concluímos, na Seção 5.1, após verificar as suposições de independência, normalidade e homocedasticidade dos erros e realizar o teste de hipóteses que, ao nível de significância de 5%, não há efeito de interação entre o fator Esporte e o bloco Gênero, assim como é desejado em um delineamento em blocos completamente aleatorizado. Posteriormente, como vimos a partir da Seção 5.2 até à Seção 5.4, a análise de diagnóstico indicou que as suposições dos erros para o modelo sem interação também foram atendidas.

Ademais, concluímos, na Seção 6, que a gordura corporal dos atletas, em média, é diferente para os três esportes analisados (corrida de barco, natação e tênis). Ainda, concluímos, na Seção 7, com 95% de confiança, que a gordura corporal dos atletas de corrida de barco é, em média, maior que a dos atletas de natação e de tênis, e também que, com 95% de confiança, não há diferença, em média, na gordura corporal dos atletas de tênis e de natação.

Referências

- [1] <https://www.med-rad.com/site/noticias/6-esportes-que-apresentam-melhores-resultados-para-perda-de-peso/> - Acesso em 08/12/2020 às 15:30.
- [2] <http://www.portalaction.com.br/analise-de-regressao/33-diagnostico-de-independencia> - Acesso em 08/12/2020 às 18:20.
- [3] ZUANETTI, Daiane Aparecida. Notas de aula de Planejamento e Análise de Experimentos (2020).
- [4] KURTNER, Michael H.; LI, William; NACHTSHEIM, Christopher J.; NETER, John - Applied Linear Statistical Models, McGraw-Hill Irwin (2004).
- [5] MONTGOMERY, Douglas C. - Design and Analysis of Experiments, Wiley (2017).

Apêndice: Códigos utilizados no R

```
# Carregando os dados
library(readxl)
DF <- read_excel("Planejamento/Trabalho2/dados_completos.xlsx")
genero=DF$V1 # bloco
esporte=DF$V2 # fator fixo
gordura_corporal=DF$Y # variável resposta
data.frame(genero,esporte,gordura_corporal)

# Sorteando os dados
# Fixaremos apenas os níveis Barco, Natação e Tênis dos Esportes
set.seed(3001) # rodar antes de todos os sorteios!
row_fem=sample(14:35,4); row_fem
swim_fem=sample(59:67,4); swim_fem
ten_fem=sample(90:96,4); ten_fem
swim_mal=sample(101:113,4); swim_mal
row_mal=sample(114:128,4); row_mal
ten_mal=199:202; ten_mal # só temos 4 amostras

dado=c(row_fem,swim_fem,ten_fem,swim_mal,row_mal,ten_mal)
y=as.numeric(gordura_corporal[dado])
genero=factor(c(rep("feminino",12),rep("masculino",12)))
esporte=factor(c(rep("barco",4),rep("natacao",4),rep("tenis",4),
  rep("natacao",4),rep("barco",4),rep("tenis",4)))
datas=data.frame(genero,esporte,y); datas

# Ajuste do modelo com interação
m<-aov(y~genero*esporte)
summary(m) # rejeita interação

# Ajuste do modelo sem interação
m1<-aov(y~genero+esporte)
summary(m1)

# Boxplot dos níveis do bloco
library(ggplot2)
ggplot(datas, aes(y = y, x = genero, fill = genero)) +
  geom_boxplot(show.legend = F, alpha = .5) +
  theme_classic(base_size = 18) +
```

```
labs(y = "Gordura Corporal", x = "Gênero",
      title = "Boxplot dos Níveis do Bloco Gênero")+
theme( plot.title = element_text(hjust = 0.5),
       axis.title = element_text(size = 11),
       axis.text = element_text(size = 13),
       title = element_text(size = 13) )

# Dotplot dos níveis do bloco
ggplot(datas, aes(x=genero, y=y, fill=genero))+
  geom_dotplot(show.legend=F, binaxis = "y", stackdir = "center",
              binwidth = 1.5, alpha = 0.8)+
labs(title = "Dotplot dos Níveis do Bloco Gênero", x = "Gênero",
     y = "Gordura Corporal") +
theme_classic()+
theme(plot.title = element_text(hjust = 0.5), axis.title =
      element_text(size = 13),
      axis.text = element_text(size = 13), title =
      element_text(size = 15))

# Boxplot dos níveis do fator
ggplot(datas, aes(y = y, x = esporte, fill = esporte)) +
  geom_boxplot(show.legend = F, alpha = .5) +
  theme_classic(base_size = 18) +
labs(y = "Gordura Corporal", x = "Esporte",
     title = "Boxplot dos Níveis do Fator Esporte")+
theme( plot.title = element_text(hjust = 0.5),
       axis.title = element_text(size = 11),
       axis.text = element_text(size = 13),
       title = element_text(size = 13) )

# Dotplot dos níveis do fator
ggplot(datas, aes(x=esporte, y=y, fill=esporte))+
  geom_dotplot(show.legend=F, binaxis = "y", stackdir = "center",
              binwidth = 1.5, alpha = 0.8)+
labs(title = "Dotplot dos Níveis do Fator Esporte", x = "Esporte",
     y = "Gordura Corporal") +
theme_classic()+
theme(plot.title = element_text(hjust = 0.5), axis.title =
      element_text(size = 13),
      axis.text = element_text(size = 13), title =
```

```

        element_text(size = 15))

# Gráfico de interação
ggplot(datas) +
  aes(x = genero, color = esporte, group = esporte, y = y) +
  stat_summary(fun = mean, geom = "point", size = 2) +
  stat_summary(fun = mean, geom = "line", size = 1)+
  labs(title = "Gráfico de interação entre fator Esporte e bloco Gênero",
        x = "Gênero", y = "Gordura Corporal") +
  theme_classic()+
  scale_colour_discrete("Esporte:") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
        element_text(size = 13),
        axis.text = element_text(size = 13), title =
        element_text(size = 13), legend.position="bottom")

# Análise de diagnóstico
res<-residuals(m1)
pred<-fitted.values(m1)
df_res <- data.frame(res, pred)

# Normalidade
(qq <- ggplot(data = df_res, aes(sample = res)) +
  stat_qq(alpha = 0.7, size = 3.5, col = "black") +
  stat_qq_line(size = 1, col = "red", lty = 1) +
  labs(title = "Gráfico Quantil-Quantil da Normal", x = "Quantis Teóricos",
        y = "Quantis Amostrais") +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5), axis.title =
        element_text(size = 13),
        axis.text = element_text(size = 13), title =
        element_text(size = 15)))

# Testes de normalidade - não rejeita em todos
library(nortest)
lillie.test(res) # Lilliefors
ks.test(res, "pnorm", mean(res), sd(res), alternative='two.sided')
shapiro.test(res) # Shapiro-Wilk
sf.test(res) # Shapiro-Francia
ad.test(res) # Anderson-Darling

```

```
# Homocedasticidade
(h <- ggplot(data = df_res) +
  geom_point(aes(x = pred, y = res), alpha = 0.7, size = 3.5, col = "black") +
  geom_line(aes(x = pred, y = 0), col = "Red", size = 1) +
  labs(title = "Gráfico Resíduos x Preditos", x = "Preditos", y = "Resíduos") +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5), axis.title =
    element_text(size = 13),
    axis.text = element_text(size = 13), title =
    element_text(size = 15)))

# Teste de homocedasticidade - não rejeita em todos
bartlett.test(split(datas$y,list(genero,esporte)))

library(PMCMRplus)
hartleyTest(split(datas$y,list(genero,esporte)))

library(car)
leveneTest(datas$y~esporte*genero, center=mean)
leveneTest(datas$y~esporte*genero, center=median)

# Independência
(ind <- ggplot(data=df_res) +
  geom_point(aes(y = res, x = 1:24), alpha = 0.7, size = 3.5, col = "black") +
  geom_hline(yintercept=0, colour="red", lwd = 1)+
  labs(title = "Gráfico Resíduos x Ordem de Coleta", x = "Ordem de Coleta",
    y = "Resíduos") +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5), axis.title =
    element_text(size = 13),
    axis.text = element_text(size = 13), title =
    element_text(size = 15)))

# Comparações múltiplas para o fator Esporte - Método de Tukey
(compar <- TukeyHSD(x=m1, conf.level=0.95))
```