**Universidade Estadual de Campinas**
Departamento de Estatística

**ME731 - Métodos em Análise Multivariada**

Professor Aluísio de Souza Pinheiro

# Regressão Multivariada - Aplicação em dados de Valorant

Luiz Felipe de Oliveira Barbosa Nunes - 255403

**Campinas - SP**
**Novembro de 2024**

# 1 Introdução

A análise de dados multivariados requer, em diversos momentos, o uso de técnicas estatísticas que permitam a modelagem simultânea de múltiplas variáveis dependentes. Entre essas técnicas, a regressão multivariada se destaca por permitir o estudo das relações entre variáveis explicativas e variáveis de resposta de maneira integrada, proporcionando uma compreensão mais abrangente e interdependente dos fatores que influenciam os resultados de interesse.

Neste trabalho, aplicaremos o modelo de regressão multivariada para analisar como diferentes métricas de desempenho em jogos de tiro em primeira pessoa, como Valorant, influenciam variáveis de interesse. Em Valorant, os jogadores são classificados em quatro classes principais: Duelistas, Controladores, Sentinelas e Iniciadores, cada uma com funções estratégicas distintas no jogo. O jogo consiste em rodadas em que a equipe atacante tenta plantar e defender uma bomba ("Spike"), enquanto a equipe defensora busca desarmá-la ou eliminar todos os adversários para vencer a rodada.

As variáveis independentes selecionadas incluem estatísticas de desempenho, como proporção de abates/-mortes (K:D), taxa de acertos na cabeça (HS%), entre outras métricas. As variáveis dependentes analisadas incluem a pontuação média de combate (ACS), o dano médio por rodada (ADR), e outras métricas que refletem a eficácia geral do jogador no jogo.

Os scripts, códigos e o conjunto de dados utilizados neste trabalho estão disponíveis no repositório GitHub: https://github.com/LuizNunes2020/AnaliseMultivariadaME731_Trab2. Para a obtenção dos dados diretamente do site https://www.vlr.gg/stats, foi utilizada uma API (Application Programming Interface), disponível no mesmo repositório, que possibilita a extração e manipulação dos dados. A extração foi realizada no dia 2/11.

# 2 Metodologia

## 2.1 Regressão Multivariada

A regressão multivariada é uma generalização da regressão linear que permite modelar a relação entre múltiplas variáveis de resposta $Y_1, Y_2, \ldots, Y_m$ e um conjunto de variáveis preditoras $Z_1, Z_2, \ldots, Z_r$. Diferente da regressão múltipla, em que há uma única variável resposta, a regressão multivariada considera que cada resposta depende das mesmas variáveis preditoras, mas cada uma possui seu próprio conjunto de coeficientes.

Nesta seção, consideramos o problema de modelar a relação entre $m$ respostas $Y_1, Y_2, \ldots, Y_m$ e um único conjunto de variáveis preditoras $z_1, z_2, \ldots, z_r$. Cada resposta é assumida como seguindo seu próprio modelo de regressão, de modo que:

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11} z_1 + \cdots + \beta_{r1} z_r + \varepsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12} z_1 + \cdots + \beta_{r2} z_r + \varepsilon_2 \\ &\vdots \\ Y_m &= \beta_{0m} + \beta_{1m} z_1 + \cdots + \beta_{rm} z_r + \varepsilon_m \end{aligned}$$

O termo de erro $\varepsilon' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m]$ tem $E(\varepsilon) = 0$ e $\mathrm{Var}(\varepsilon) = \Sigma$. Assim, os termos de erro associados a diferentes respostas podem ser correlacionados.

Além disso, para estabelecer uma notação que se conforma ao modelo clássico de regressão linear, deixe $[z_{j0}, z_{j1}, \ldots, z_{jr}]$ denotar os valores das variáveis preditoras para o $j$-ésimo teste, deixe $Y_j' = [Y_{j1}, Y_{j2}, \ldots, Y_{jm}]$ ser as respostas, e deixe $\varepsilon_j' = [\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jm}]$ ser os erros. Em notação matricial, a matriz de projeto

$$Z = \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix}_{n \times (r+1)}$$

é a mesma que para o modelo de regressão de resposta única. As outras quantidades matriciais têm contrapartes multivariadas, sendo:

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{bmatrix}_{n \times m} = \left[ Y_{(1)} | Y_{(2)} | \ldots | Y_{(m)} \right]$$

$$\beta = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{bmatrix}_{(r+1) \times m} = \left[ \beta_{(1)} | \beta_{(2)} | \ldots | \beta_{(m)} \right]$$

$$\varepsilon = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix}_{n \times m} = \left[ \varepsilon_{(1)} | \varepsilon_{(2)} | \ldots | \varepsilon_{(m)} \right]$$

O modelo de regressão linear multivariada é

$$Y = Z\beta + \varepsilon$$

com $E(\varepsilon_{(ij)}) = 0$ e $\text{Cov}(\varepsilon_{(i)}, \varepsilon_{(j)}) = \sigma_{ik} I \quad i, k = 1, 2, \ldots, m$.

As $m$ observações no $j$-ésimo teste têm matriz de covariância $\Sigma = \{\sigma_{ik}\}$, mas as observações de testes diferentes são não correlacionadas. Aqui $\beta$ e $\sigma_{ik}$ são parâmetros desconhecidos; a matriz de projeto $Z$ tem $j$-ésima linha $[z_{j0}, z_{j1}, \ldots, z_{jr}]$.

Assim, a $i$-ésima resposta $Y_{(i)}$ segue o modelo de regressão linear

$$Y_{(i)} = Z\beta_{(i)} + \varepsilon_{(i)}, \quad i = 1, 2, \ldots, m$$

com $\text{Cov}(\varepsilon_{(i)}) = \sigma_{ii} I$. No entanto, os erros para diferentes respostas no mesmo teste podem ser correlacionados.

Dada as observações $Y$ e os valores das variáveis preditoras $Z$ com posto total, determinamos as estimativas de mínimos quadrados $\hat{\beta}_{(i)}$ exclusivamente das observações $Y_{(i)}$ na $i$-ésima resposta. Em conformidade com a solução de resposta única, tomamos

$$\hat{\beta} = (Z'Z)^{-1}Z'Y$$

Coletando essas estimativas de mínimos quadrados univariadas, obtemos

$$\hat{\beta} = [\hat{\beta}_{(1)}|\hat{\beta}_{(2)}|\ldots|\hat{\beta}_{(m)}] = (Z'Z)^{-1}Z'\left[Y_{(1)}|Y_{(2)}|\ldots|Y_{(m)}\right]$$

ou

$$\hat{\beta} = (Z'Z)^{-1}Z'Y$$

Para qualquer escolha de parâmetros $B = [b_{(1)}|b_{(2)}|\ldots|b_{(m)}]$, a matriz de erros é $Y - ZB$. A soma de quadrados e produtos cruzados do erro é

$$(Y - ZB)'(Y - ZB) = \begin{bmatrix} (Y_{(1)} - Zb_{(1)})'(Y_{(1)} - Zb_{(1)}) & \ldots & (Y_{(1)} - Zb_{(1)})'(Y_{(m)} - Zb_{(m)}) \\ \vdots & \ddots & \vdots \\ (Y_{(m)} - Zb_{(m)})'(Y_{(1)} - Zb_{(1)}) & \ldots & (Y_{(m)} - Zb_{(m)})'(Y_{(m)} - Zb_{(m)}) \end{bmatrix}$$

A seleção dos parâmetros $b_{(i)} = \hat{\beta}_{(i)}$ minimiza a i-ésima soma diagonal dos quadrados

$$(Y_{(i)} - Zb_{(i)})'(Y_{(i)} - Zb_{(i)}).$$

Consequentemente, $tr[(Y - ZB)'(Y - ZB)]$ é minimizado pela escolha $B = \hat{B}$. Além disso, a variância generalizada $|(Y - ZB)'(Y - ZB)|$ é minimizada pelas estimativas de mínimos quadrados $\hat{B}$.

Usando as estimativas de mínimos quadrados $\hat{B}$, podemos formar as matrizes de valores previstos e resíduos:

$$\hat{Y} = Z\hat{B} = Z(Z'Z)^{-1}Z'Y$$

$$\hat{\epsilon} = Y - \hat{Y} = [I - Z(Z'Z)^{-1}Z']Y$$

As condições de ortogonalidade entre os resíduos, valores previstos e colunas de $Z$, que valem para a regressão linear clássica, também valem para a regressão múltipla multivariada. Elas seguem de $Z'[I - Z(Z'Z)^{-1}Z'] = Z' - Z' = 0$. Especificamente,

$$Z'\hat{\epsilon} = Z'[I - Z(Z'Z)^{-1}Z']Y = 0$$

de modo que os resíduos $\hat{\epsilon}_{(i)}$ são perpendiculares às colunas de $Z$. Além disso,

$$\hat{Y}'\hat{\epsilon} = \hat{B}'Z'[I - Z(Z'Z)^{-1}Z']Y = 0$$

confirmando que os valores previstos $\hat{Y}_{(i)}$ são perpendiculares a todos os vetores residuais $\hat{\epsilon}_{(i)}$. Como $Y = \hat{Y} + \hat{\epsilon}$,

$$Y'Y = (\hat{Y} + \hat{\epsilon})'(\hat{Y} + \hat{\epsilon}) = \hat{Y}'\hat{Y} + \hat{\epsilon}'\hat{\epsilon} + 0 + 0'$$

ou

$$Y'Y = \hat{Y}'\hat{Y} + \hat{\epsilon}'\hat{\epsilon}$$

onde:

(soma total de quadrados e produtos cruzados) =

(soma prevista de quadrados e produtos cruzados) + (soma de erro residual de quadrados e produtos cruzados)

A soma residual dos quadrados e produtos cruzados também pode ser escrita como:

$$\hat{\epsilon}'\hat{\epsilon} = Y'Y - \hat{Y}'\hat{Y} = Y'Y - \hat{\beta}'Z'Z\hat{\beta}$$

Dividindo cada entrada $\hat{\epsilon}'_{(i)}\hat{\epsilon}_{(k)}/(n-r-1)$, obtemos o estimador não viciado de $\Sigma$. Finalmente,

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_{(i)}, \hat{\epsilon}_{(k)}) &= E[(Z'Z)^{-1}Z'\hat{\epsilon}_{(k)}\hat{\epsilon}'_{(i)}(I - Z(Z'Z)^{-1}Z')] \\ &= (Z'Z)^{-1}Z'\sigma_{ik}(I - Z(Z'Z)^{-1}Z') \\ &= \sigma_{ik}(Z'Z)^{-1}.\end{aligned}$$

Assim, cada elemento de $\hat{\beta}$ é não correlacionado com cada elemento de $\hat{\epsilon}$.

## 2.2 Testes de Razão de Verossimilhança para Parâmetros de Regressão

O análogo multi-resposta da hipótese de que as respostas não dependem das variáveis $z_{q+1}, z_{q+2}, \ldots, z_r$ é descrito pela seguinte hipótese estatística:

$$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_r = 0 \quad \text{ou} \quad H_0 : \beta_{(2)} = 0$$

onde o vetor de coeficientes $\beta$ é particionado como:

$$\beta = \begin{bmatrix}\beta_{(1)} \\ \beta_{(2)}\end{bmatrix}$$

Neste contexto, $\beta_{(1)}$ corresponde aos coeficientes associados às variáveis preditoras mantidas no modelo, enquanto $\beta_{(2)}$ representa os coeficientes das variáveis cuja influência nas variáveis de resposta estamos testando, para verificar se podem ser excluídas do modelo.

Definindo $Z = \begin{bmatrix}Z_1 & Z_2\end{bmatrix}$, podemos escrever o modelo geral como

$$E(Y) = Z\beta = \begin{bmatrix}Z_1 & Z_2\end{bmatrix}\begin{bmatrix}\beta_{(1)} \\ \beta_{(2)}\end{bmatrix} = Z_1\beta_{(1)} + Z_2\beta_{(2)}$$

Sob a hipótese nula $H_0 : \beta_{(2)} = 0$, o modelo se reduz a $Y = Z_1\beta_{(1)} + \varepsilon$, e o teste de razão de verossimilhança para $H_0$ é baseado nas somas extras de quadrados e produtos:

$$(Y - Z_1\hat{\beta}_{(1)})'(Y - Z_1\hat{\beta}_{(1)}) - (Y - Z\hat{\beta})'(Y - Z\hat{\beta})$$

$$= n(\hat{\Sigma}_1 - \hat{\Sigma})$$

onde $\hat{\beta}_{(1)} = (Z_1'Z_1)^{-1}Z_1'Y$ e $\hat{\Sigma}_1 = \frac{1}{n}(Y - Z_1\hat{\beta}_{(1)})'(Y - Z_1\hat{\beta}_{(1)})$.

De acordo com o **resultado 1**:

Seja o modelo de regressão múltipla multivariada com $Z$ de posto completo, $\text{rank}(Z) = r+1$, $n \geqslant (r+1)+m$, e suponha que os erros $\varepsilon$ tenham uma distribuição normal. Então, o estimador de máxima verossimilhança de $\beta$ é dado por

$$\hat{\beta} = (Z'Z)^{-1}Z'Y$$

e $\hat{\beta}$ tem uma distribuição normal com

$$\mathbb{E}(\hat{\beta}) = \beta \quad \text{e} \quad \text{Cov}(\hat{\beta}_{(i)}, \hat{\beta}_{(k)}) = \sigma_{ik}(Z'Z)^{-1}.$$

Além disso, $\hat{\beta}$ é independente do estimador de máxima verossimilhança do estimador positivo definido $\Sigma$, dado por

$$\hat{\Sigma} = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n}(Y - Z\hat{\beta})'(Y - Z\hat{\beta}),$$

onde $n\hat{\Sigma}$ é distribuído como $W_{p,n-r-1}(\Sigma)$.

A função de verossimilhança maximizada $L(\hat{\mu}, \hat{\Sigma})$ é dada por

$$L(\hat{\mu}, \hat{\Sigma}) = (2\pi)^{-mn/2}|\hat{\Sigma}|^{-n/2}e^{-mn/2}.$$

Com base nesse resultado, a razão de verossimilhança, $\Lambda$, pode ser expressa em termos de variâncias generalizadas como:

$$\Lambda = \frac{\max\limits_{\beta_{(1)},\Sigma} L(\beta_{(1)}, \Sigma)}{\max\limits_{\beta,\Sigma} L(\beta, \Sigma)} = \frac{|\hat{\Sigma}|^{n/2}}{|\hat{\Sigma}_1|^{n/2}} \tag{7-38}$$

De forma equivalente, a estatística lambda de Wilks pode ser usada para o teste:

$$\Lambda^{2n} = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}$$

## 2.3 Outras Estatísticas de Teste Multivariadas

Testes além do teste de razão de verossimilhança foram propostos para testar $H_0 : \beta_{(2)} = 0$ no modelo de regressão múltipla multivariada.

Seja $E$ o erro $p \times p$, ou matriz de soma de quadrados e produtos de resíduos, com

$$E = n\hat{\Sigma}$$

resultante do ajuste do modelo completo. A matriz $p \times p$ da hipótese, ou matriz de soma de quadrados e produtos extra, é

$$H = n(\hat{\Sigma}_1 - \hat{\Sigma})$$

As estatísticas podem ser definidas em termos de $E$ e $H$ diretamente, ou em termos dos autovalores não nulos $\eta_1 \geqslant \eta_2 \geqslant \cdots \geqslant \eta_s$ de $HE^{-1}$, onde $s = \min(p, f-q)$. Equivalentemente, elas são as raízes de $|\hat{\Sigma}_1 - \hat{\Sigma} - \eta I| = 0$. As definições são

- **Lambda de Wilks**: $\Lambda = \prod_{i=1}^{s} \frac{1}{1+\eta_i} = \frac{|E|}{|E+H|}$
- **Traço de Pillai**: $\sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = \text{tr}[(H+E)^{-1}]$

- **Traço de Hotelling-Lawley**: $\sum_{i=1}^{s} \eta_i = \text{tr}[HE^{-1}]$
- **Maior Raiz de Roy**: $\frac{\eta_1}{1+\eta_1}$

O teste de Roy seleciona o vetor de coeficiente $v$ de modo que a estatística $F$ univariada baseada em $a'Y$ tenha seu valor máximo possível. Quando vários dos autovalores $\eta_i$ são moderadamente grandes, o teste de Roy irá apresentar um desempenho ruim em relação aos outros três. Estudos de simulação sugerem que seu poder será melhor quando houver apenas um autovalor grande.

Tabelas de gráficos de valores críticos estão disponíveis para o teste de Roy. Lambda de Wilks, maior raiz de Roy e o traço de Hotelling-Lawley são testes quase equivalentes para tamanhos de amostra grandes.

Se houver uma grande discrepância nos valores-P para os quatro testes, os autovalores e vetores próprios podem levar a uma interpretação.

## 2.4 Previsões em Regressões Multivariadas

Suponha que o modelo $Y = Z\beta + \varepsilon$, com erros normais $\varepsilon$, tenha sido ajustado e verificado quanto à adequação. Se o modelo for adequado, ele pode ser empregado para fins preditivos.

Um problema comum é prever as respostas médias correspondentes a valores fixos $z_0$ das variáveis preditoras. Inferências sobre as respostas médias podem ser feitas utilizando a teoria de distribuição do **resultado 1**.

A partir deste resultado, determinamos que

$$\hat{\beta}z_0 \text{ é distribuído como } N_m(\beta' z_0, z_0'(Z'Z)^{-1}z_0\Sigma)$$

e

$$n\hat{\Sigma} \text{ é distribuído independentemente como } W_{n-r-1}(\Sigma).$$

O valor desconhecido da função de regressão em $z_0$ é $\beta' z_0$. Assim, podemos definir a estatística $T^2$ como:

$$T^2 = \left(\hat{\beta}z_0 - \beta' z_0\right)' \left[\frac{n}{n-r-1}\hat{\Sigma}\right]^{-1} \left(\hat{\beta}z_0 - \beta' z_0\right) (z_0'(Z'Z)^{-1}z_0),$$

e o elipsoide de confiança de $100(1-\alpha)\%$ para $\beta' z_0$ é fornecido pela desigualdade

$$\left(\hat{\beta}z_0 - \beta' z_0\right)' \left(\frac{n}{n-r-1}\hat{\Sigma}\right)^{-1} \left(\hat{\beta}z_0 - \beta' z_0\right) \leqslant z_0'(Z'Z)^{-1}z_0 \left[\frac{m(n-r-1)}{n-r-m}\right] F_{m,n-r-m}(\alpha),$$

onde $F_{m,n-r-m}(\alpha)$ é o percentil superior de $100\alpha\%$ de uma distribuição $F$ com $m$ e $n-r-m$ graus de liberdade.

Os intervalos de confiança simultâneos de $100(1-\alpha)\%$ para $E(Y_i) = z_0\beta_{(i)}$ são

$$z_0\hat{\beta}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m}\right) F_{m,n-r-m}(\alpha) z_0'(Z'Z)^{-1}z_0 \frac{n}{n-r-1}\hat{\sigma}_{ii}}, \quad i = 1, 2, \ldots, m,$$

onde $\hat{\beta}_{(i)}$ é a i-ésima coluna de $\hat{\beta}$ e $\hat{\sigma}_{ii}$ é o i-ésimo elemento diagonal de $\hat{\Sigma}$.

O segundo problema de previsão envolve prever novas respostas $Y_0 = \beta' z_0 + e_0$ em $z_0$, onde $e_0$ é independente de $\varepsilon$. Neste caso,

$$Y_0 - \hat{\beta}z_0 = (\beta - \hat{\beta})' z_0 + e_0 \text{ é distribuído como } N_m(0, (1 + z_0'(Z'Z)^{-1}z_0)\Sigma),$$

independentemente de $n\hat{\Sigma}$. Portanto, o elipsoide de previsão $100(1-\alpha)\%$ para $Y_0$ é

$$(Y_0 - \hat{\beta}z_0)' \left(\frac{n}{n-r-1}\hat{\Sigma}\right)^{-1} (Y_0 - \hat{\beta}z_0) \leqslant (1 + z_0'(Z'Z)^{-1}z_0) \left[\frac{m(n-r-1)}{n-r-m}\right] F_{m,n-r-m}(\alpha).$$

Os intervalos de previsão simultâneos de $100(1-\alpha)\%$ para as respostas individuais $Y_{0i}$ são

$$z_0\hat{\beta}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m}\right) F_{m,n-r-m}(\alpha)(1 + z_0'(Z'Z)^{-1}z_0) \frac{n}{n-r-1}\hat{\sigma}_{ii}}, \quad i = 1, 2, \ldots, m,$$

onde $\hat{\beta}_{(i)}$, $\hat{\sigma}_{ii}$, e $F_{m,n-r-m}(\alpha)$ são as mesmas quantidades que aparecem nos intervalos de confiança acima. Observa-se que os intervalos de previsão para os valores *reais* das variáveis de resposta são mais largos do que os intervalos para os valores *esperados*, devido ao erro aleatório $e_{0i}$.

## 2.5 Análise Multivariada de Variância (MANOVA)

A Análise Multivariada de Variância (MANOVA) é utilizada para avaliar a significância dos preditores no modelo de regressão multivariada. No nosso caso, a tabela MANOVA é configurada para o modelo restrito, onde a versão matricial da soma de quadrados dos resíduos, $S_\varepsilon^*$, no modelo restrito é dada por

$$S_\varepsilon^* = (Y - \hat{\mu} - \hat{C}X)(Y - \hat{\mu} - \hat{C}X)^\top,$$

ou, expandindo e reorganizando os termos,

$$S_\varepsilon^* = (Y - \hat{\mu}X)(Y - \hat{\mu}X)^\top + (\hat{C} - \hat{\mu})XX^\top(\hat{C} - \hat{\mu})^\top,$$

onde o primeiro termo representa a soma de quadrados dos resíduos para o modelo irrestrito e o segundo termo é a fonte adicional de variação, $S_h = S_e - S_\varepsilon^*$, devido à remoção das restrições. Os termos de produtos cruzados desaparecem devido à ortogonalidade entre os resíduos e os valores ajustados.

Para o modelo irrestrito, a soma de quadrados da regressão, $S_{\text{reg}}$, é dada por

$$S_{\text{reg}} = \hat{C}XX^\top\hat{C}^\top.$$

A soma de quadrados da regressão para o modelo restrito é denotada por $S_{\text{reg}}^*$.

Utilizando essa configuração, podemos resumir os resultados em uma tabela MANOVA, na qual comparamos as somas de quadrados dos modelos restrito e irrestrito para avaliar a significância dos preditores. A diferença entre esses modelos fornece uma estatística de teste para a hipótese de que as variáveis preditoras não têm efeito significativo sobre as variáveis de resposta.

A matriz $S_h$ pode ser expressa mais explicitamente como:

$$S_h = K^\top(KK^\top)^{-1}(K\hat{C}L - \Gamma)(L^\top(XX^\top)^{-1}L)^{-1}(K\hat{C}L - \Gamma)^\top(KK^\top)^{-1}K.$$

Substituindo e expandindo, obtemos a expectativa de $S_h$ como:

$$\mathbb{E}(S_h) = D(K\hat{C}L - \Gamma)(L^\top(XX^\top)^{-1}L)^{-1}(K\hat{C}L - \Gamma)^\top D^\top + F \cdot \mathbb{E}(GEG^\top) \cdot F^\top,$$

onde $D = K^\top(KK^\top)^{-1}$, $F = DK$, e $G = X^\top(XX^\top)^{-1}L$. Essas matrizes ajudam a projetar o espaço das variáveis de resposta dentro do subespaço relevante das variáveis preditoras, permitindo que a MANOVA capte a influência multivariada dos preditores.

A estatística lambda de Wilks pode ser usada como critério para avaliar a significância dos preditores. A estatística é dada por:

$$\Lambda = \frac{|\hat{S}_{\text{res}}|}{|\hat{S}_{\text{res}} + \hat{S}_{\text{reg}}|},$$

onde $\hat{S}_{\text{res}}$ representa a soma de quadrados e produtos cruzados dos resíduos e $\hat{S}_{\text{reg}}$ a soma de quadrados e produtos cruzados da regressão. Quando o valor de $\Lambda$ é baixo, indica que há uma diferença significativa entre os grupos em relação às variáveis de resposta, sugerindo um efeito significativo das variáveis preditoras no modelo multivariado.

## 2.6 Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é uma técnica de análise multivariada que objetiva realizar uma projeção ortogonal da matriz de dados em um espaço de menor dimensão, maximizando a variabilidade representada. Segundo Izenman (2008), seja $S = \frac{1}{n}(X - \bar{X})^T(X - \bar{X})$ a matriz de covariância estimada das

colunas de $X$, onde $\bar{X}$ é a matriz cujas colunas correspondem às médias de cada coluna de $X$. Os autovalores e autovetores dessa matriz são, respectivamente, $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \cdots \geqslant \hat{\lambda}_p \geqslant 0$ e $(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_p)$, onde $\hat{\lambda}_i$ está associado ao autovetor $\hat{v}_i$, para $i = 1, 2, \ldots, p$.

A melhor reconstrução de $X$ com posto $t < p$ é dada por:

$$\hat{X}^{(t)} = \bar{X} + \sum_{i=1}^{t} \hat{v}_i \hat{v}_i^T (X - \bar{X}),$$

onde o escore da $j$-ésima componente principal de $X$ é estimado por:

$$\hat{\psi}_j = \hat{v}_j^T (X - \bar{X}),$$

e a variância da $j$-ésima componente principal é estimada por $\hat{\lambda}_j$. Uma medida da qualidade da projeção, em termos de variabilidade explicada, é dada pela proporção da variabilidade representada pelas $t$ primeiras componentes principais, ou seja:

$$\frac{\sum_{j=1}^{t} \hat{\lambda}_j}{\sum_{j=1}^{p} \hat{\lambda}_j}.$$

# 3 Aplicação no conjunto de dados

O banco de dados utilizado possui um total de 21 variáveis e $n = 554$ observações. Na regressão multivariada, foram consideradas quatro variáveis dependentes:

- **kill_deaths**: razão entre o número de eliminações (kills) e mortes.
- **average_damage_per_round**: dano médio causado por rodada.
- **average_combat_score**: pontuação média de combate do jogador por rodada.
- **rating**: pontuação geral do jogador, refletindo sua performance geral.

As variáveis independentes, excluindo os identificadores como "player"e "org", ficando 15 variáveis independentes são:

- **rounds_played**: número de rodadas jogadas.
- **kill_assists_survived_traded**: porcentagem de rodadas em que o jogador conseguiu uma eliminação, assistência, sobreviveu ou foi trocado.
- **kills_per_round**: número médio de eliminações por rodada.
- **assists_per_round**: número médio de assistências por rodada.
- **first_kills_per_round**: número médio de primeiros abates por rodada.
- **first_deaths_per_round**: número médio de primeiras mortes por rodada.
- **headshot_percentage**: porcentagem geral de tiros na cabeça (*headshots*).
- **clutch_success_percentage**: porcentagem de sucesso em situações de *clutch*.
- **clutch_wins**: número total de vitórias em situações de *clutch*.
- **max_kills**: número máximo de eliminações em uma rodada.
- **kills**: número total de eliminações.
- **deaths**: número total de mortes.
- **assists**: número total de assistências.
- **first_kills**: número total de primeiros abates.
- **first_deaths**: número total de primeiras mortes.

**Observação:** Uma situação de *clutch* refere-se a momentos críticos em que uma jogada decisiva pode definir o resultado final da partida. Esses momentos frequentemente ocorrem sob alta pressão, como nas fases finais de uma rodada ou quando o placar está equilibrado.

Em primeiro lugar, carreguei o banco de dados e preparei os dados transformando colunas de porcentagem em valores numéricos, ajustando a variável `clutch_wins` e removendo valores ausentes. Em seguida, visualizei a distribuição das variáveis numéricas usando histogramas e boxplots. Além disso, realizei análises bivariadas para explorar as relações entre as variáveis dependentes e independentes.

Para a modelagem, apliquei Análise de Componentes Principais (PCA) nas variáveis independentes, buscando reduzir a dimensionalidade e minimizar a multicolinearidade. As seis primeiras componentes principais, que explicam 93% da variância, foram usadas em uma regressão multivariada para prever as variáveis dependentes, e utilizei MANOVA para avaliar a significância geral dos preditores. Finalizei o processo aplicando diagnósticos de resíduos, testes de normalidade e homocedasticidade e verifiquei a linearidade entre as componentes principais e as variáveis dependentes, garantindo a adequação do modelo ajustado.

## 3.1 Análise Descritiva

A Tabela 1 apresenta as estatísticas descritivas para cada variável do conjunto de dados. Destacam-se alguns pontos importantes: a variável *rounds_played* possui um valor máximo de 1005, significativamente acima da média de 367,1, indicando a presença de jogadores com extensa participação em partidas. Observa-se que *rating* varia de 0,61 a 1,48, com uma média muito próxima de 1, sugerindo que a maioria dos jogadores tem desempenho consistente, com poucos se destacando ou ficando muito abaixo dessa média. Adicionalmente, *average_combat_score* e *kills_per_round* também apresentam uma ampla variação, refletindo a diversidade de habilidades e estilos de jogo dos participantes.

Tabela 1: Análise descritiva do conjunto de dados

| **Variável** | **Mínimo** | **1º Quartil** | **Mediana** | **Média** | **3º Quartil** | **Máximo** |
|---|---|---|---|---|---|---|
| rounds_played | 201.0 | 245.2 | 312.0 | 367.1 | 448.0 | 1005.0 |
| rating | 0.6100 | 0.9100 | 0.9900 | 0.9922 | 1.0800 | 1.4800 |
| average_combat_score | 124.0 | 177.6 | 195.9 | 198.5 | 218.3 | 307.2 |
| kill_deaths | 0.52 | 0.87 | 0.98 | 0.99 | 1.10 | 1.80 |
| kill_assists_survived_traded | 0.5500 | 0.6700 | 0.7000 | 0.7032 | 0.7400 | 0.8500 |
| average_damage_per_round | 88.3 | 119.3 | 129.8 | 131.1 | 143.2 | 194.6 |
| kills_per_round | 0.4300 | 0.6200 | 0.6900 | 0.6985 | 0.7700 | 1.1000 |
| assists_per_round | 0.100 | 0.210 | 0.280 | 0.288 | 0.360 | 0.610 |
| first_kills_per_round | 0.02000 | 0.06000 | 0.08000 | 0.09776 | 0.12000 | 0.29000 |
| first_deaths_per_round | 0.0300 | 0.0700 | 0.0900 | 0.1007 | 0.1200 | 0.2800 |
| headshot_percentage | 0.1600 | 0.2500 | 0.2900 | 0.2889 | 0.3200 | 0.4600 |
| clutch_success_percentage | 0.0300 | 0.1000 | 0.1400 | 0.1459 | 0.1800 | 0.3500 |
| clutch_wins | 0.02564 | 0.09859 | 0.14286 | 0.14579 | 0.18265 | 0.35000 |
| max_kills | 11.00 | 21.00 | 24.00 | 24.43 | 27.00 | 43.00 |
| kills | 97.0 | 168.0 | 217.0 | 257.5 | 312.0 | 784.0 |
| deaths | 101.0 | 176.2 | 230.0 | 260.9 | 315.0 | 726.0 |
| assists | 22.0 | 64.0 | 88.0 | 106.5 | 133.5 | 393.0 |
| first_kills | 5.00 | 19.00 | 29.00 | 36.25 | 45.00 | 199.00 |
| first_deaths | 9.00 | 21.00 | 31.00 | 36.96 | 46.00 | 161.00 |

A Figura 1 mostra os histogramas das variáveis numéricas. Variáveis como *rating* e *headshot_percentage* exibem distribuições com uma concentração próxima à média, indicando uma distribuição relativamente simétrica. Por outro lado, variáveis como *rounds_played* e *kills_per_round* têm distribuições assimétricas à direita, com uma concentração de jogadores na faixa de valores mais baixos e alguns casos extremos que representam jogadores com alto desempenho.

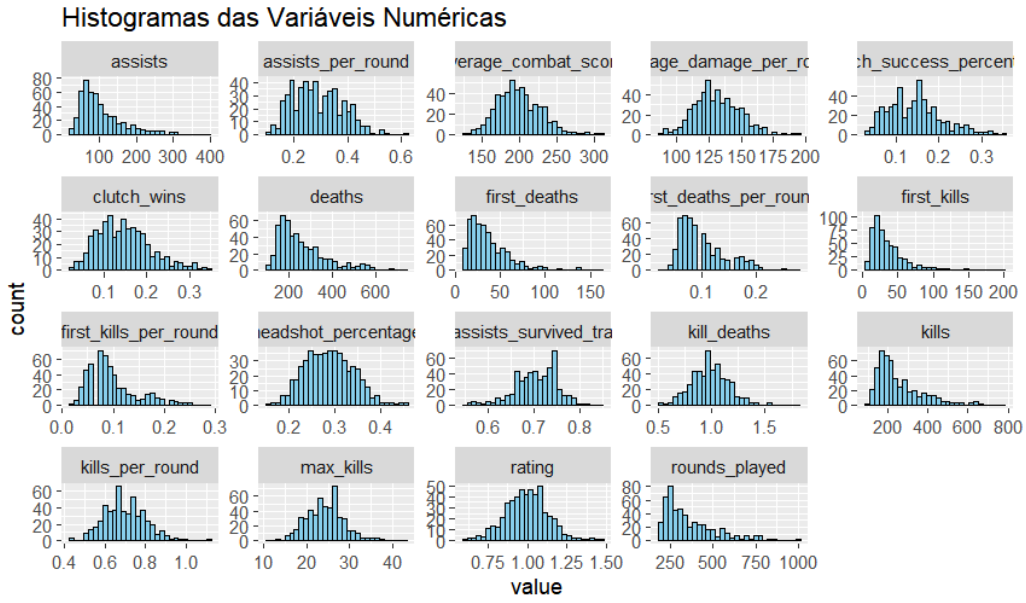



Figura 1: Histograma das variáveis numéricas

Na Figura 2, são apresentados os boxplots das variáveis numéricas, destacando visualmente a presença de outliers. Observa-se que as variáveis *rounds_played* e *kills* possuem valores muito superiores aos seus quartis superiores. Outras variáveis como rating tem quase nenhuma variabilidade, concentrando seus valores quase em um único ponto.

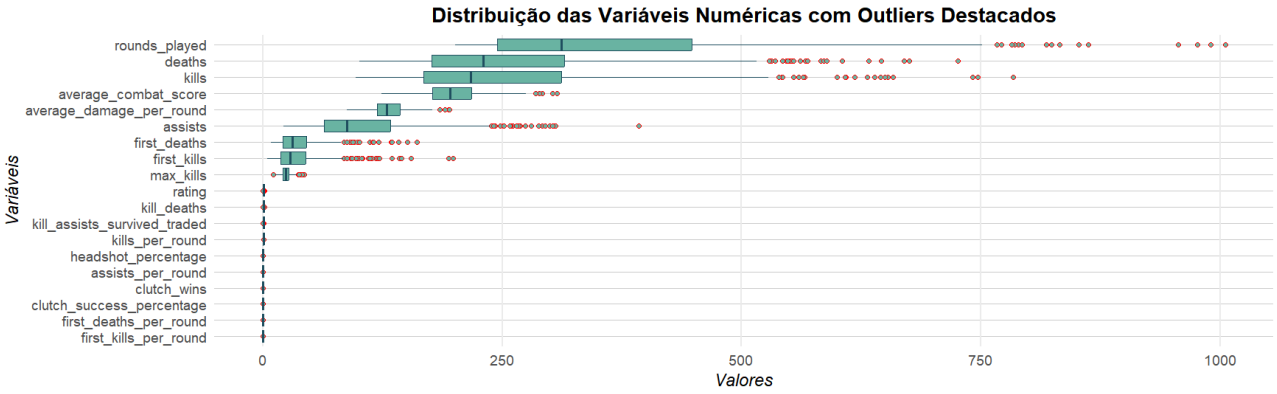
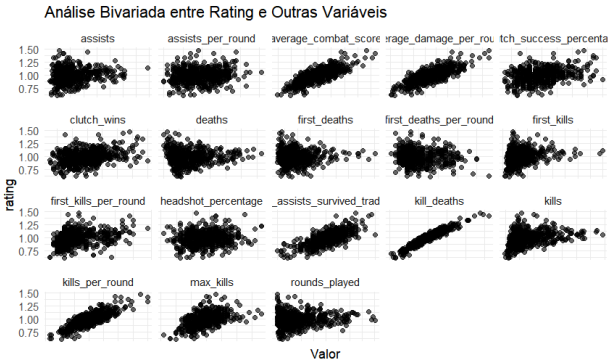
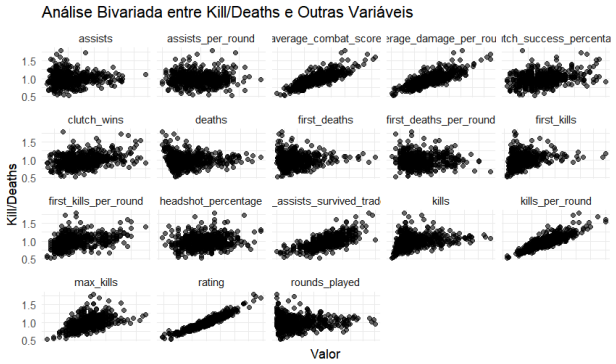



Figura 2: Boxplot das variáveis numéricas

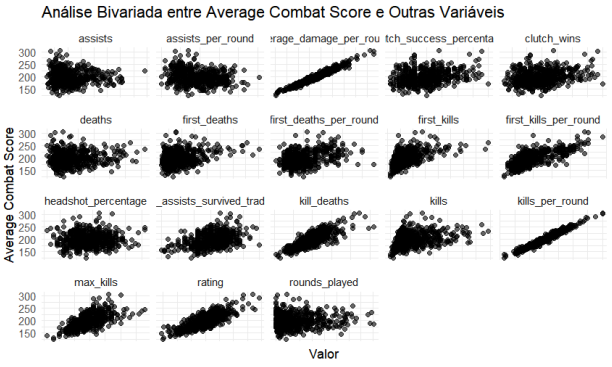
Pelos gráficos 3, observa-se que *rating* apresenta uma relação positiva com variáveis como *kills_per_round* e *average_combat_score*, sugerindo que maiores pontuações de combate e número de eliminações por rodada contribuem para o aumento do rating do jogador. Similarmente verifica-se que *average_combat_score* está fortemente correlacionado com *average_damage_per_round*, indicando que o dano médio é um fator importante na pontuação de combate.



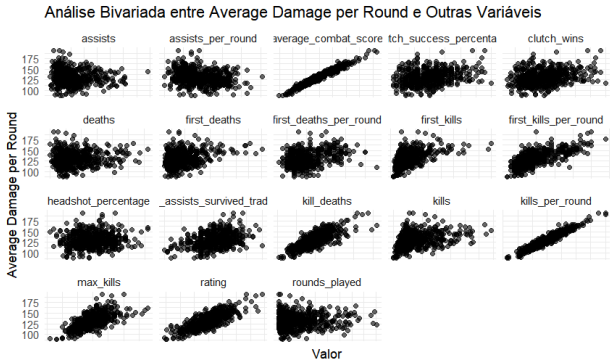

(a) Análise Bivariada - Rating




(b) Análise Bivariada - Kill/Deaths




(c) Análise Bivariada - Average Combat Score




(d) Análise Bivariada - Average Damage per Round

Figura 3: Análise Bivariada das Variáveis Dependentes

A Tabela 2 apresenta a matriz de covariância dos resíduos entre as variáveis de resposta. Notamos uma correlação positiva marcante entre *average_damage_per_round* e *average_combat_score*, sugerindo que um aumento em uma dessas variáveis tende a estar associado ao aumento da outra. Essa correlação implica que ambas as variáveis medem aspectos complementares do desempenho.

Tabela 2: Matriz de Covariância dos Resíduos

| | **kill_deaths** | **average_damage_per_round** | **average_combat_score** | **rating** |
|---|---|---|---|---|
| **kill_deaths** | 0.0067 | 0.1735 | 0.3015 | 0.0036 |
| **average_damage_per_round** | 0.1735 | 49.1693 | 61.5227 | 0.2095 |
| **average_combat_score** | 0.3015 | 61.5227 | 101.5243 | 0.3110 |
| **rating** | 0.0036 | 0.2095 | 0.3110 | 0.0030 |

A Tabela 2 exibe os valores do Fator de Inflação da Variância (VIF) para as variáveis independentes, indicando a presença de multicolinearidade. Destaca-se o VIF elevado para *clutch_success_percentage* e *clutch_wins*, sugerindo que essas variáveis estão altamente correlacionadas com outras métricas (multicolinearidade).

Devido à alta taxa de multicolinearidade, optei por realizar uma Análise de Componentes Principais (PCA) para reduzir a dimensionalidade do conjunto de dados. A PCA permite transformar as variáveis independentes em um conjunto de componentes principais não correlacionadas, minimizando a redundância entre as variáveis originais e melhorando a robustez da modelagem. A Tabela 3 apresenta a importância de cada componente principal, com as seis primeiras explicando juntas aproximadamente 93% da variância dos dados. Para visualizar a contribuição de cada componente, foi realizado o *Scree Plot*, ilustrado na Figura 4.

Tabela 3: Importância das Componentes Principais

| Componente | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Desvio-padrão | 2.3274 | 1.7793 | 1.5570 | 1.1710 | 1.0409 | 0.7756 | 0.6692 | 0.4296 | 0.3936 | 0.3236 | 0.1430 | 0.1160 | 0.0849 | 0.0517 | 0.0303 |
| Proporção de Variância | 0.3611 | 0.2111 | 0.1616 | 0.0914 | 0.0722 | 0.0401 | 0.0299 | 0.0123 | 0.0103 | 0.0070 | 0.0014 | 0.0009 | 0.0005 | 0.0002 | 0.0001 |
| Proporção Cumulativa | 0.3611 | 0.5722 | 0.7338 | 0.8252 | 0.8975 | 0.9375 | 0.9674 | 0.9797 | 0.9900 | 0.9970 | 0.9984 | 0.9993 | 0.9998 | 0.9999 | 1.0000 |

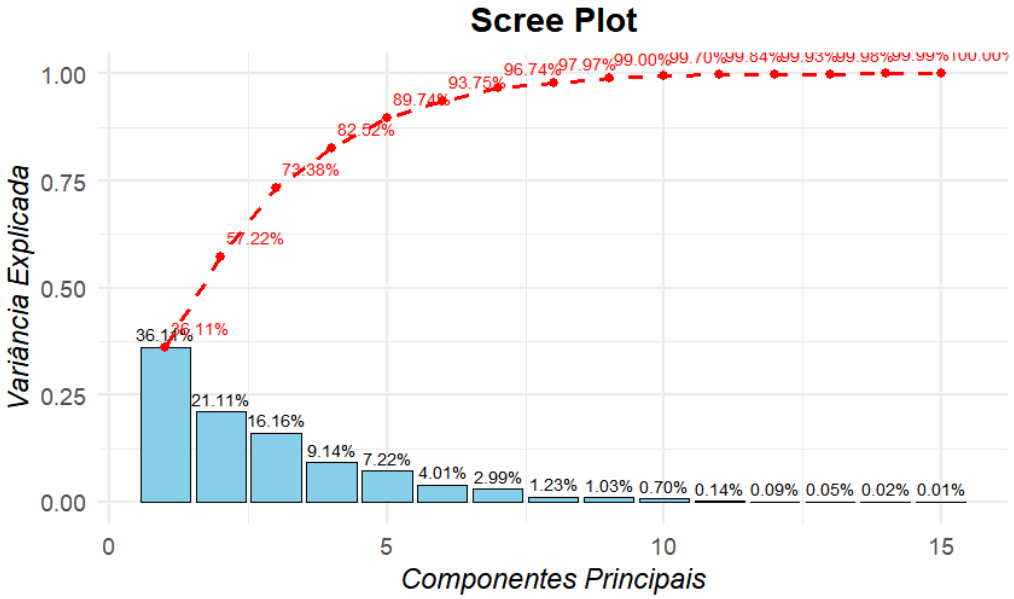



Figura 4: Scree Plot da PCA

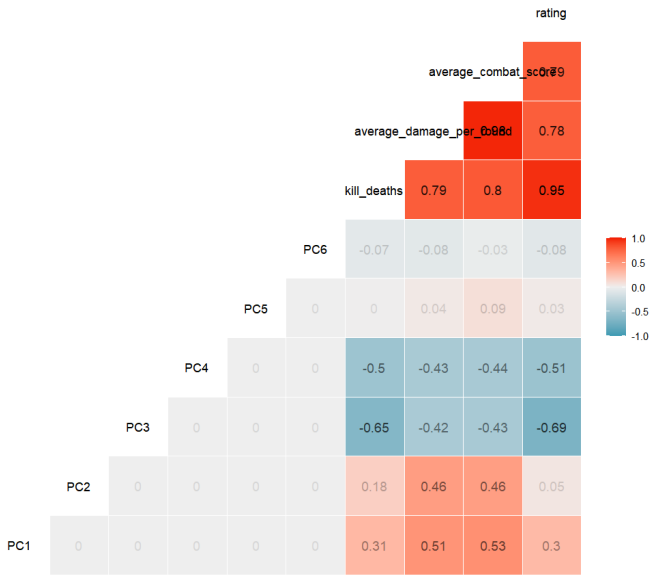



Figura 5: Análise de correlação entre as componentes principais e as variáveis dependentes

Na Figura 5, destaca-se a forte correlação positiva entre *kill_deaths* e *average_combat_score* com a primeira componente principal (PC1), indicando que essa componente captura uma parte significativa das variações nessas métricas. Além disso, a variável *rating* mostra uma correlação considerável com a terceira componente principal (PC3), sugerindo que essa componente pode estar associada a fatores que influenciam diretamente a avaliação geral do jogador. Notamos ainda que a componente PC4 apresenta uma correlação negativa com *average_combat_score* e *average_damage_per_round*, sugerindo que essa componente captura uma dimensão oposta ao desempenho ofensivo direto.

# 4 Resultados

As Tabelas a seguir apresentam os resultados da regressão multivariada aplicada para cada variável dependente (*kill_deaths*, *average_damage_per_round*, *average_combat_score* e *rating*) utilizando as componentes principais selecionadas.

Tabela 4: Regressão Multivariada para a Variável Dependente *kill_deaths*

| **Coeficiente** | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** |
|---|---|---|---|---|
| Intercept | 0.9899805 | 0.0036268 | 272.963 | $< 2e-16$ *** |
| PC1 | 0.0245684 | 0.0015598 | 15.751 | $< 2e-16$ *** |
| PC2 | 0.0182864 | 0.0020403 | 8.963 | $< 2e-16$ *** |
| PC3 | -0.0761811 | 0.0023316 | -32.673 | $< 2e-16$ *** |
| PC4 | -0.0778807 | 0.0031002 | -25.121 | $< 2e-16$ *** |
| PC5 | 0.0005913 | 0.0034875 | 0.170 | 0.865441 |
| PC6 | -0.0163318 | 0.0046809 | -3.489 | 0.000527 ** |

Legenda: Significância dos coeficientes: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$, . indica $p < 0.1$.

Na Tabela 1, observamos que as componentes principais **PC1**, **PC2**, **PC3** e **PC4** têm coeficientes altamente significativos ($p < 0.001$), indicando que essas componentes são fundamentais para explicar a variabilidade na variável *kill_deaths*. Em particular, **PC3**, **PC4** e **PC6** apresentam coeficientes negativos.

Tabela 5: Regressão Multivariada para a Variável Dependente *average_damage_per_round*

| **Coeficiente** | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** |
|---|---|---|---|---|
| Intercept | 131.0817 | 0.3111 | 421.330 | $< 2e-16$ *** |
| PC1 | 3.8580 | 0.1338 | 28.833 | $< 2e-16$ *** |
| PC2 | 4.5314 | 0.1750 | 25.891 | $< 2e-16$ *** |
| PC3 | -4.7354 | 0.2000 | -23.676 | $< 2e-16$ *** |
| PC4 | -6.3913 | 0.2659 | -24.033 | $< 2e-16$ *** |
| PC5 | 0.6314 | 0.2992 | 2.110 | 0.0353 * |
| PC6 | -1.7008 | 0.4015 | -4.236 | $2.71e-05$ *** |

Na Tabela 2, para a variável *average_damage_per_round*, todas as componentes principais, exceto **PC5**, apresentam alta significância ($p < 0.01$). O coeficiente positivo de **PC1** e **PC2** sugere que essas componentes aumentam o *average_damage_per_round*, enquanto **PC3** e **PC4** possuem coeficientes negativos, indicando uma relação inversa com o dano médio por rodada.

Tabela 6: Regressão Multivariada para a Variável Dependente *average_combat_score*

| **Coeficiente** | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** |
|---|---|---|---|---|
| Intercept | 198.5329 | 0.4471 | 444.093 | $< 2e-16$ *** |
| PC1 | 6.5433 | 0.1923 | 34.031 | $< 2e-16$ *** |
| PC2 | 7.3862 | 0.2515 | 29.370 | $< 2e-16$ *** |
| PC3 | -7.9746 | 0.2874 | -27.747 | $< 2e-16$ *** |
| PC4 | -10.6813 | 0.3821 | -27.951 | $< 2e-16$ *** |
| PC5 | 2.4668 | 0.4299 | 5.738 | $1.65e-08$ *** |
| PC6 | -1.0710 | 0.5770 | -1.856 | 0.064 . |

A Tabela 3 mostra que **PC1**, **PC2**, **PC3** e **PC4** são altamente significativas para prever *average_combat_score*, com coeficientes elevados que refletem uma forte influência no desempenho médio de combate. A componente **PC5** também é significativa, mas em menor grau, enquanto **PC6** não apresenta significância estatística.

Tabela 7: Regressão Multivariada para a Variável Dependente *rating*

| Coeficiente | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 0.992237 | 0.002426 | 409.029 | $< 2e-16$ *** |
| PC1 | 0.017629 | 0.001043 | 16.897 | $< 2e-16$ *** |
| PC2 | 0.004189 | 0.001365 | 3.070 | 0.00226 ** |
| PC3 | -0.060857 | 0.001560 | -39.023 | $< 2e-16$ *** |
| PC4 | -0.059776 | 0.002074 | -28.827 | $< 2e-16$ *** |
| PC5 | 0.003701 | 0.002333 | 1.587 | 0.11321 |
| PC6 | -0.014577 | 0.003131 | -4.656 | $4.12e-06$ *** |

Na Tabela 4, os resultados para *rating* mostram que **PC1**, **PC3**, e **PC4** têm um impacto significativo ($p < 0.001$), destacando-se como principais influenciadores no cálculo do rating geral do jogador. **PC2** e **PC6** também são significativas, mas em menor grau, enquanto **PC5** não é estatisticamente significativa.

## 4.1 Diagnósticos

A Tabela 8 apresenta os resultados do teste de Wilks' Lambda para avaliar a significância das componentes principais no modelo. Notamos que as componentes principais **PC1** a **PC4** apresentam valores de Wilks' Lambda baixos e estatisticamente significativos ($p < 2.2 \times 10^{-16}$), indicando que essas componentes são altamente relevantes para explicar a variabilidade nas variáveis dependentes. Em particular, **PC3** apresenta o menor valor de Wilks' Lambda (0.24058), sugerindo que essa componente captura uma parte significativa da variabilidade do modelo. Em contrapartida, as componentes **PC5** e **PC6** possuem valores mais elevados de Wilks' Lambda (0.87338 e 0.91741, respectivamente), com significância ainda relevante, mas menor, o que indica que elas contribuem de forma menos expressiva para o modelo.

Tabela 8: Resultados do Teste de Wilks' Lambda para as Componentes Principais

| Componente | Df | Wilks' Lambda | approx F | num Df | den Df | Pr(>F) |
|---|---|---|---|---|---|---|
| PC1 | 1 | 0.28868 | 310.46 | 4 | 504 | $< 2.2e-16$ *** |
| PC2 | 1 | 0.25600 | 366.19 | 4 | 504 | $< 2.2e-16$ *** |
| PC3 | 1 | 0.24058 | 397.73 | 4 | 504 | $< 2.2e-16$ *** |
| PC4 | 1 | 0.32186 | 265.47 | 4 | 504 | $< 2.2e-16$ *** |
| PC5 | 1 | 0.87338 | 18.27 | 4 | 504 | $5.021e-14$ *** |
| PC6 | 1 | 0.91741 | 11.34 | 4 | 504 | $8.038e-09$ *** |

Na Tabela 9, os resultados do teste de Shapiro-Wilk para normalidade dos resíduos das variáveis dependentes indicam que, para a variável *kill_deaths*, o p-valor ($2.18 \times 10^{-8}$) é significativamente menor que 0.05, sugerindo que os resíduos não seguem uma distribuição normal. No entanto, para as variáveis *average_damage_per_round*, *average_combat_score*, e *rating*, os p-valores são maiores que 0.05 (com valores de 0.08321, 0.2694 e 0.3847, respectivamente), o que indica que os resíduos destas variáveis não violam a suposição de normalidade de forma significativa.

Tabela 9: Teste de Normalidade de Shapiro-Wilk para os Resíduos de Cada Variável Dependente

| Variável | Estatística W | p-valor |
|---|---|---|
| kill_deaths | 0.97179 | 2.18e-08 |
| average_damage_per_round | 0.99485 | 0.08321 |
| average_combat_score | 0.99626 | 0.2694 |
| rating | 0.99672 | 0.3847 |

A Tabela 10 mostra os resultados do teste de homocedasticidade (ncvTest) para cada variável dependente, avaliando se os resíduos possuem variância constante. Para *kill_deaths*, o teste resultou em um p-valor extremamente baixo ($7.98 \times 10^{-13}$), indicando violação da homocedasticidade, ou seja, a variância dos resíduos não

é constante. A variável *rating* também mostra um p-valor baixo (0.00046), indicando potencial heterocedasticidade. No entanto, as variáveis *average_damage_per_round* e *average_combat_score* apresentam p-valores de 1.00000 e 0.42484, respectivamente, sugerindo que não há evidências de heterocedasticidade para esses resíduos.

Tabela 10: Teste de Homocedasticidade (ncvTest) para Cada Variável Dependente

| Variável | Estatística Qui-Quadrado | Graus de Liberdade | p-valor |
|---|---|---|---|
| kill_deaths | 51.28748 | 1 | 7.98e-13 |
| average_damage_per_round | 0.00000 | 1 | 1.00000 |
| average_combat_score | 0.63690 | 1 | 0.42484 |
| rating | 12.28361 | 1 | 0.00046 |

Os resultados destes testes de diagnóstico indicam que as suposições de normalidade e homocedasticidade são satisfeitas para a maioria das variáveis dependentes, exceto para *kill_deaths*, que apresenta desvios em ambas as suposições.

Após realizar diversas tentativas de transformação para satisfazer as suposições de normalidade e homocedasticidade nos resíduos, os resultados ainda não foram satisfatórios. As transformações aplicadas incluíram a transformação logarítmica (*log*), Box-Cox, raiz quadrada (*sqrt*), recíproca (*1/x*), raiz cúbica (*cbrt*), exponencial inversa ($1/e^x$), e arco seno (*arcsinh*).

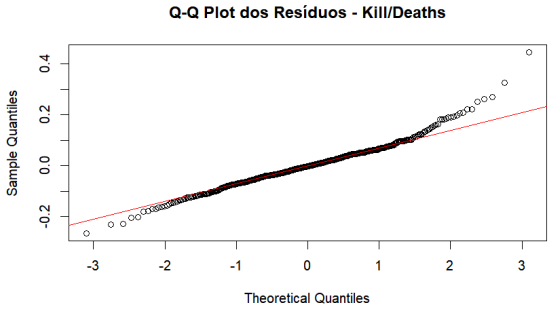



Figura 6: QQ-plot dos Resíduos - Kill/Deaths

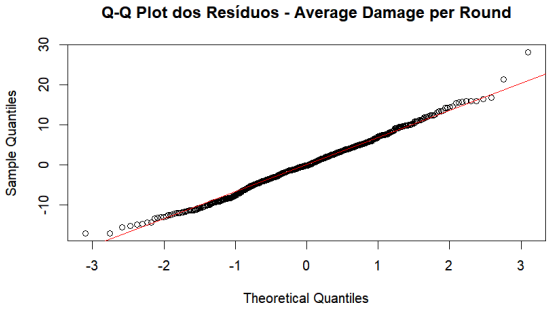



Figura 7: QQ-plot dos Resíduos - Average Damage per Round

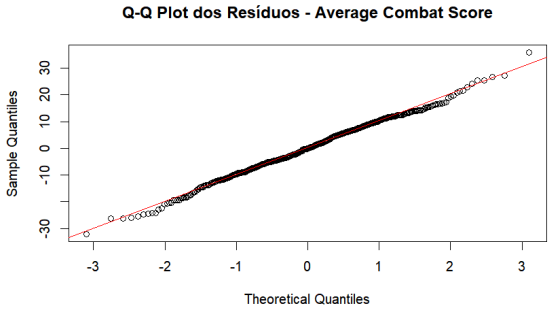



Figura 8: QQ-plot dos Resíduos - Average Combat Score

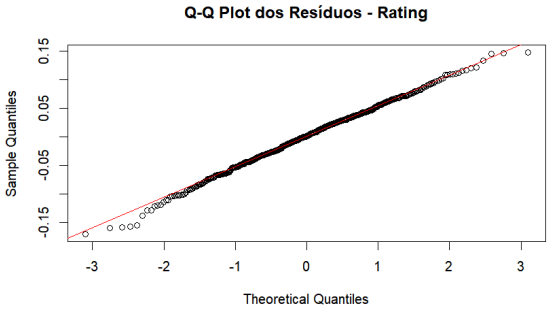



Figura 9: QQ-plot dos Resíduos - Rating

Figura 10: QQ-plots dos Resíduos das Variáveis Dependentes

Na Figura 10, são apresentados os QQ-plots dos resíduos para cada variável dependente do modelo. Esses gráficos permitem avaliar a aderência dos resíduos à normalidade. O QQ-plot dos resíduos de *Kill/Deaths* (Figura 6) mostra uma discrepância considerável em relação à linha de normalidade, especialmente nas caudas, indicando uma possível distribuição assimétrica ou presença de outliers. Já o QQ-plot dos resíduos de *Average Damage per Round* (Figura 7) mostra uma aproximação maior com a linha de normalidade, com pequenas

variações nas extremidades. Para a variável *Average Combat Score* (Figura 8), também se observa um leve desvio nas caudas, porém menos pronunciado que em *Kill/Deaths*. Por fim, o QQ-plot dos resíduos de *Rating* (Figura 9) mostra uma aderência muito próxima à linha de normalidade, indicando que os resíduos dessa variável seguem uma distribuição normal mais adequada. Esses resultados indicam que, a suposição de normalidade é violada principalmente para *Kill/Deaths*.

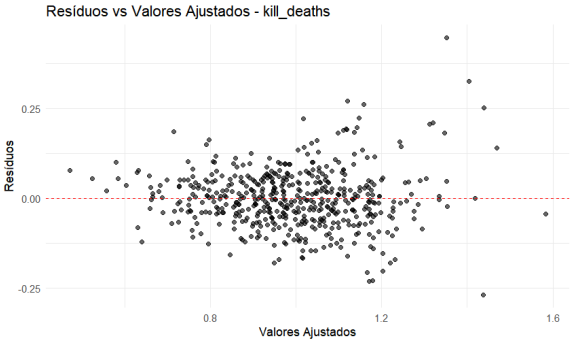



Figura 11: Resíduos vs Valores Ajustados - Kill/Deaths

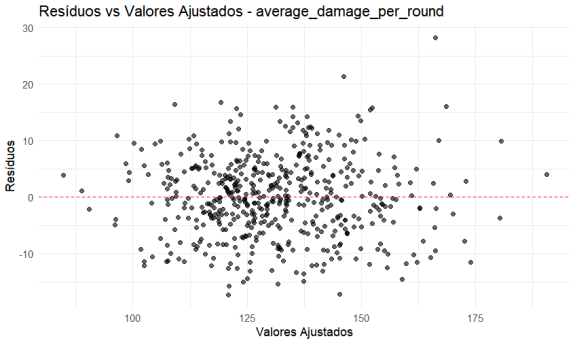



Figura 12: Resíduos vs Valores Ajustados - Average Damage per Round

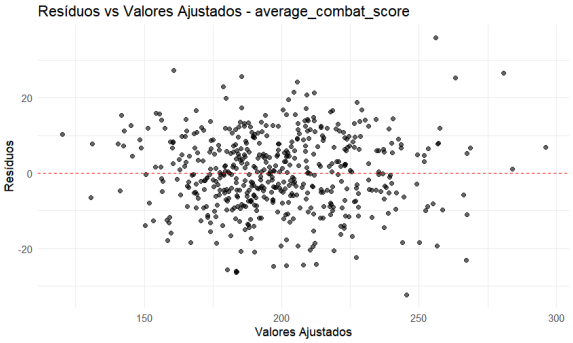



Figura 13: Resíduos vs Valores Ajustados - Average Combat Score

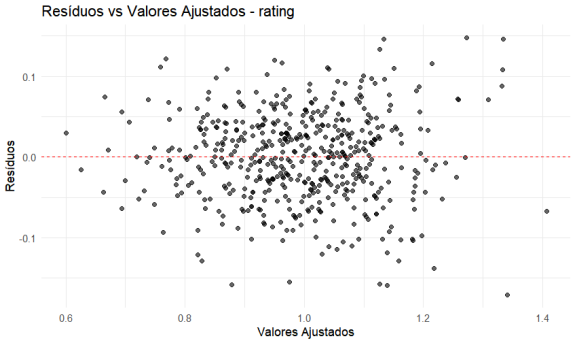



Figura 14: Resíduos vs Valores Ajustados - Rating

Figura 15: Gráficos de Resíduos vs Valores Ajustados para as Variáveis Dependentes

Na Figura 15, os gráficos de resíduos versus valores ajustados mostram a distribuição dos resíduos para cada variável dependente. No caso de *Kill/Deaths*, observa-se uma leve heterocedasticidade, com concentração de pontos e alguns valores extremos. Para *Average Damage per Round* e *Average Combat Score*, os resíduos estão razoavelmente dispersos, mas com variações nos extremos, sugerindo não-constância de variância. Já para *Rating*, a dispersão é mais uniforme, indicando que a suposição de homocedasticidade está mais próxima de ser atendida. Esses resultados sugerem que o modelo está melhor ajustado para *Rating* em comparação com *Kill/Deaths*.

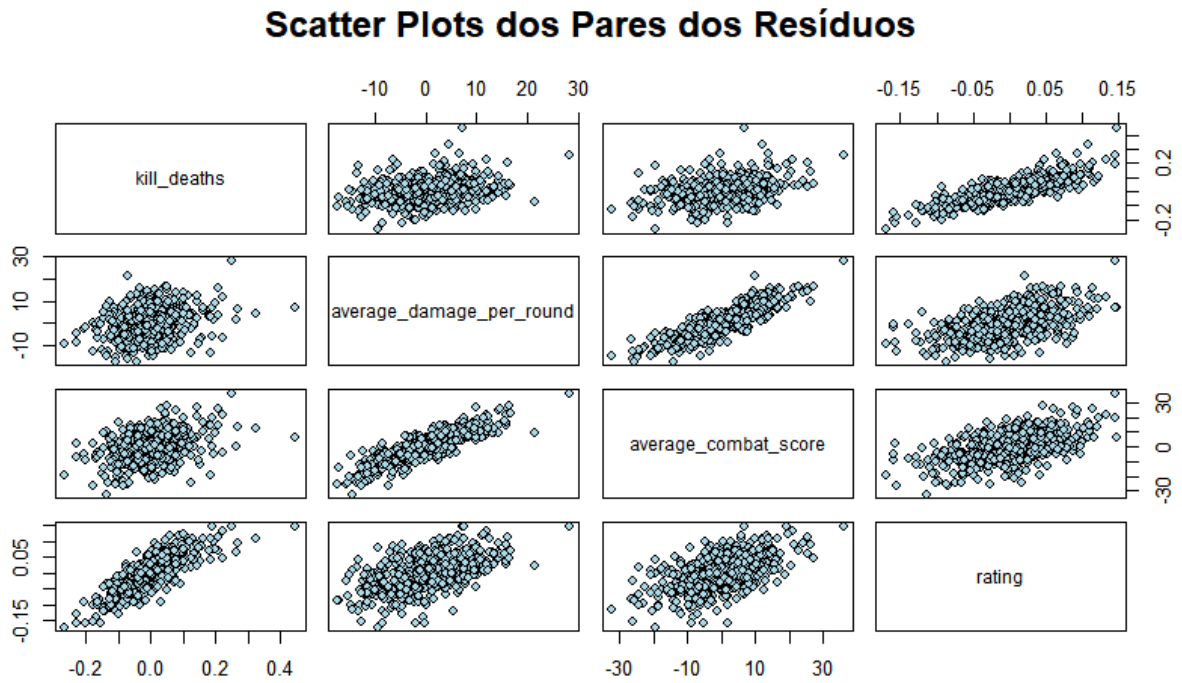



Figura 16: Scatter Plot dos pares dos resíduos

Para verificar a linearidade entre as principais componentes e as variáveis de resposta ajustadas, foram gerados gráficos que mostram a relação entre as componentes principais PC1 e PC2 com as variáveis de resposta ajustadas.

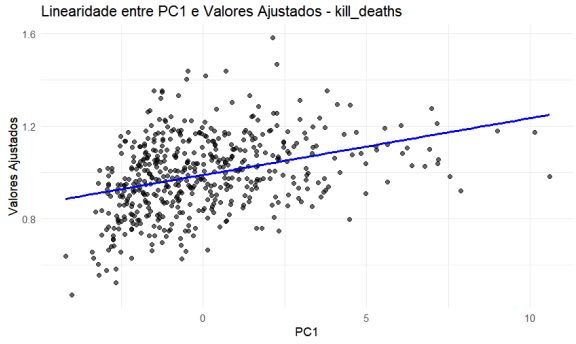



Figura 17: Linearidade entre PC1 e *kill_deaths*

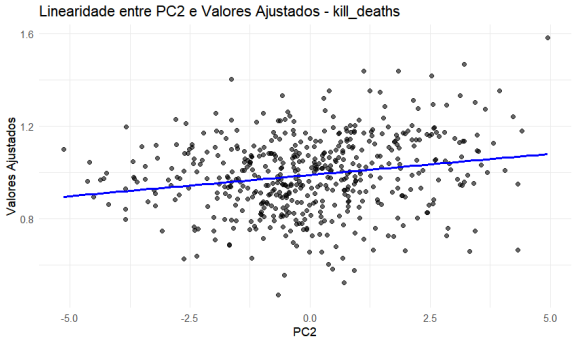



Figura 18: Linearidade entre PC2 e *kill_deaths*

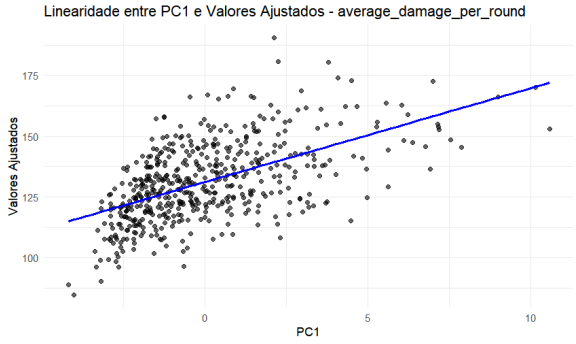



Figura 19: Linearidade entre PC1 e *average_damage_per_round*

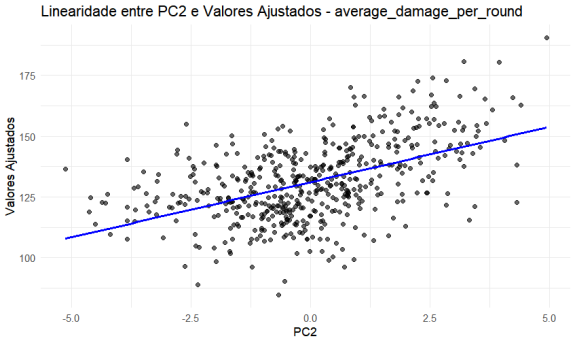



Figura 20: Linearidade entre PC2 e *average_damage_per_round*

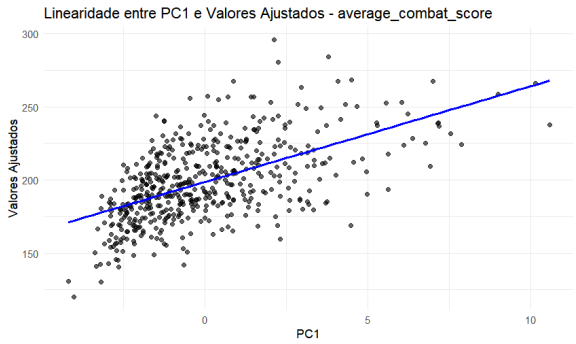



Figura 21: Linearidade entre PC1 e *average_combat_score*

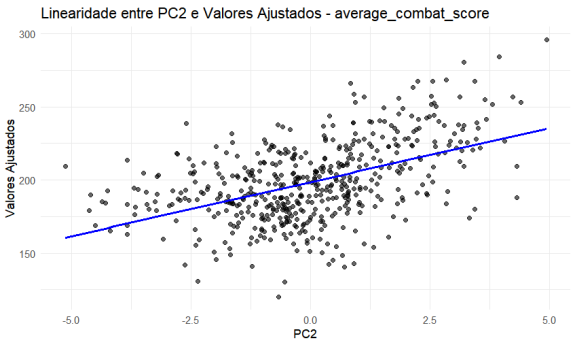



Figura 22: Linearidade entre PC2 e *average_combat_score*

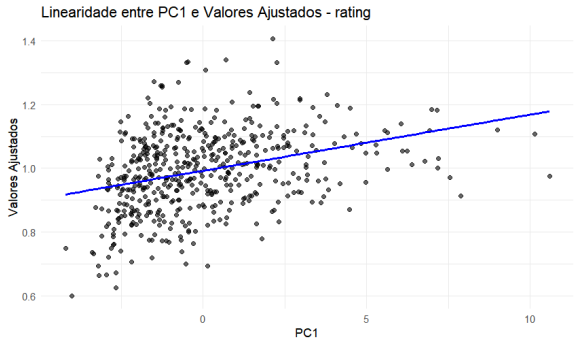



Figura 23: Linearidade entre PC1 e *rating*

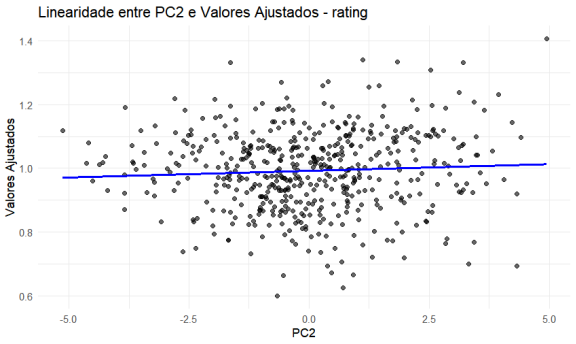



Figura 24: Linearidade entre PC2 e *rating*

## 4.2 Curiosidade

Para cada variável de resposta, calculei intervalos de confiança com base nos preditores do modelo, proporcionando uma estimativa do possível intervalo onde os valores reais podem se situar. Abaixo, apresentei os intervalos de confiança das primeiras seis observações:

Tabela 11: Intervalos de Confiança para as Previsões

| Observação | kill_deaths | average_damage_per_round | average_combat_score | rating |
|---|---|---|---|---|
| 1 | 1.4384 | 166.2272 | 255.9685 | 1.3344 |
| 2 | 1.4039 | 150.2031 | 231.5086 | 1.3324 |
| 3 | 1.4695 | 180.6362 | 280.6533 | 1.3319 |
| 4 | 1.3525 | 157.1018 | 243.5422 | 1.2721 |
| 5 | 1.4191 | 166.8898 | 257.4365 | 1.3090 |
| 6 | 1.5837 | 190.5501 | 296.0176 | 1.4069 |

# 5 Conclusão

Os resultados indicaram que as componentes principais **PC1**, **PC2**, **PC3** e **PC4** possuem grande relevância para explicar a variabilidade nas variáveis dependentes, conforme indicado pelos testes de Wilks' Lambda e pelas análises de significância dos coeficientes de regressão. A variável *rating*, em particular, mostrou uma relação mais consistente com as componentes principais, sugerindo que o modelo está melhor ajustado para essa métrica em comparação com outras, como *kill_deaths*, que apresentou desvios nas suposições de normalidade e homocedasticidade dos resíduos. Para ilustrar a contribuição de cada variável nas componentes principais, incluí na Figura 25 um gráfico de barras com as cargas das variáveis em cada componente principal

Apesar das diversas tentativas de transformação dos resíduos, as suposições de normalidade e homocedasticidade não foram completamente atendidas para todas as variáveis de resposta, indicando que o modelo pode ser aprimorado. Possíveis direções futuras incluem a aplicação de métodos não lineares ou técnicas de regularização para melhorar o ajuste e a robustez do modelo.

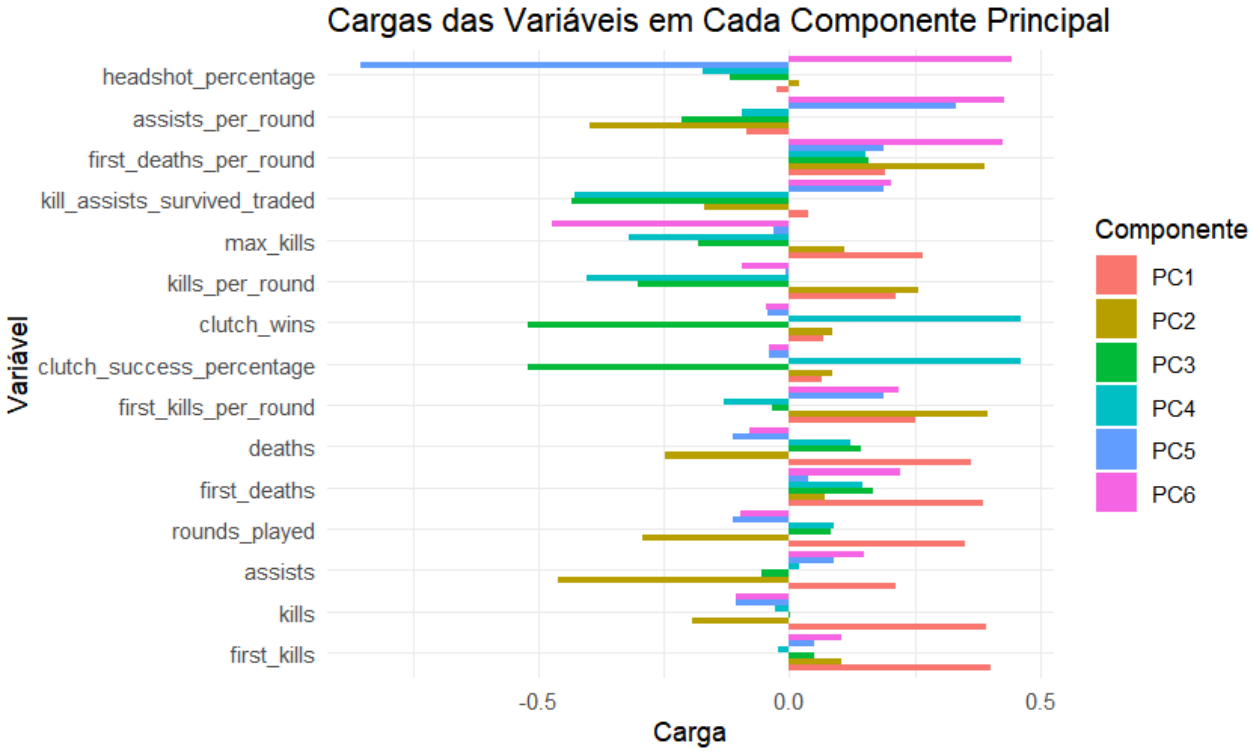



Figura 25: Gráfico de barras para as cargas das variáveis em cada componente principal

# 6 Discussão Crítica

## 6.1 Modelo de Regressão Multivariada

O modelo de regressão multivariada é uma ferramenta valiosa para analisar a relação entre várias variáveis independentes e múltiplas variáveis dependentes simultaneamente. Ele permite capturar interações entre preditores, enriquecendo a análise e proporcionando uma visão mais completa.

Apesar das vantagens, o modelo enfrenta desafios. A complexidade aumenta com o número de variáveis, exigindo maior capacidade computacional. A multicolinearidade entre preditores pode tornar as estimativas instáveis, dificultando a interpretação dos efeitos individuais e reduzindo a precisão das inferências.

Ademais, o modelo assume linearidade, homocedasticidade e normalidade dos resíduos. A violação dessas suposições compromete a validade dos resultados, podendo exigir transformações de dados ou métodos alternativos. Outliers também representam um problema, pois podem distorcer as estimativas, sendo fundamental seu tratamento adequado.

# Referências

[1] Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. 6th edition. Prentice Hall. Disponível em: https://www.webpages.uidaho.edu/~stevel/519/Applied%20Multivariate%20Statistical%20Analysis%20by%20Johnson%20and%20Wichern.pdf

[2] University of Virginia Library (2020). *Getting started with Multivariate Multiple Regression*. Disponível em: https://library.virginia.edu/data/articles/getting-started-with-multivariate-multiple-regression

[3] R Core Team (2023). *Classical Multivariate Regression*. Disponível em: https://cloud.r-project.org/web/packages/rrr/vignettes/rrr.html

[4] Artes, R., & Barroso, L. P. (2005). *Métodos Multivariados de Análise em Estatística*. São Paulo: Editora XYZ.

[5] A. Julian Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Science  Business Media, 2008.

[6] https://www.vlr.gg/stats. Acesso em: 02 nov. 2024.