ISSN: 2317-0840

Modelos de regressão para o valor da produção de arroz no Rio Grande do Sul

Iverson Rodrigues Custódio[1], Ana Lúcia Souza Silva Mateus[2†], Luciane Flores Jacobi[3]

[1]*Bacharel em Estatística; Departamento de Estatística; Universidade Federal de Santa Maria; Santa Maria – RS, Brasil.*
[2]*Professora Adjunta no Departamento de Estatística; Universidade Federal de Santa Maria; Santa Maria – RS, Brasil.*
[3]*Professora Titular no Departamento de Estatística, Universidade Federal de Santa Maria; Santa Maria – RS, Brasil.*

Resumo: *O arroz é o segundo grão mais consumido no mundo. No Brasil, o Rio Grande do Sul (RS) é o maior produtor de arroz em casca. Dessa forma, o objetivo deste estudo é avaliar quais são os fatores que influenciam o valor da produção deste cereal. Para explicar o valor da produção de arroz foram utilizadas as variáveis de produção obtidas no site do Departamento de Economia e Estatística e as variáveis meteorológicas retiradas no site Brazilian Daily Weather Gridded Data ,referentes ao ano de 2019. As variáveis analisadas apresentam uma distribuição não normal e heterocedástica justificando o uso da técnica de regressão linear generalizada. A seleção das variáveis para os modelos foi feita através do método stepwise. Os modelos obtidos mostraram que a média anual de radiação solar e a média anual de precipitação de chuva influenciam o valor da produção de arroz no RS, assim como em outros trabalhos. No entanto, esses modelos podem não ser adequados para capturar a relação não linear entre o valor de produção de grãos e variáveis como clima e produção.*

Palavras-chave: *Distribuição gama; modelo linear generalizado; modelo linear; arroz.*

Regression models for the value of rice production in Rio Grande do Sul

Abstract: *Rice is the second most consumed grain in the world. In Brazil, Rio Grande do Sul (RS) is the largest producer of paddy rice. Therefore, the objective of this study is to evaluate the factors that influence the production value of this cereal. To explain the value of rice production, production variables obtained from the Department of Economics and Statistics website and derived variables taken from the Brazilian Daily Weather Gridded Data website were used, referring to the year 2019. The variables found present a non-normal and heteroscedastic distribution, justifying the use of the generalized linear regression technique. The selection of variables for the models was done using the stepwise method. The models found demonstrated that the annual average of solar radiation and the annual average of rainfall transfer influence the value of rice production in RS, as well as in other studies. However, these models may not be suitable to capture the non-linear relationship between grain production value and variables such as climate and production.*

Keywords: *Gamma distribution; generalized linear model; linear model; rice.*

[†]Autora correspondente: ana.mateus@ufsm.br

Manuscrito recebido em: 29/07/2024
Manuscrito revisado em: 20/10/2024
Manuscrito aceito em: 29/10/2024

Introdução

O arroz é um dos grãos mais consumidos no Brasil e um alimento indispensável nas cestas básicas. Além disso, sua ampla produção desperta interesse em explorar seu potencial. Entre as unidades da federação, o Rio Grande do Sul (RS) é o maior produtor de arroz em casca do Brasil. Segundo a Pesquisa Agrícola Municipal do IBGE, o RS registrou no período 2018-2020 uma produção de 7.775.850 toneladas em média do grão ATLAS (2022). Logo torna-se importante o estudo de variáveis que expliquem o fato de o território gaúcho ser uma potência na produção desta cultivar.

Wickramasinghe *et al.* (2020) destacam que a variação climática é uma das questões mais impactantes para a cultura do arroz, gerando efeitos significativos sobre a produção. Esses mesmos autores utilizaram diversos métodos para compreender as relações entre os fatores climáticos e a colheita. Além disso, variáveis econômicas apresentam relevância ao estudar o valor de produção de plantações. Ribeiro (2021) apresenta algumas variáveis econômicas ao estudar a produção de erva-mate, levantando a questão se essas variáveis também interferem na produção de arroz.

Modelos de regressão linear são muito utilizados para previsão de variáveis em muitas áreas de conhecimentos como para estimar o número de óbitos por acidente vascular cerebral no estado do Rio Grande do Sul (Silva *et al.*, 2019), o número de casos de AIDS para cada estado da Região Norte do Brasil (Da Silva e Jacobi, 2020), assim como para encontrar variáveis que tem influência sobre o valor de imóveis residenciais (Mateus *et al*., 2019).

Dessa forma, embora seja amplamente utilizada, apresenta certas limitações, pois exige que a variável resposta apresente normalidade, o que nem sempre ocorre. Quando essa condição não é atendida, um método frequentemente utilizado é a regressão linear generalizada, que abrange várias distribuições da família exponencial, sendo uma expansão da regressão linear (McCULLAGH; NELDER, 1989).

A regressão robusta com métodos dos momentos foi utilizada por Nugrahani et al. (2021) para prever a quantidade de arroz produzida. O modelo mostrou que quanto maior a área colhida e a população, maior será a quantidade de arroz produzida, e que o aumento de chuva e pragas de plantas interfere negativamente na quantidade de produção de arroz.

Em outro estudo em uma província no noroeste de Sri Lanka, foram usadas algumas técnicas de regressão (regressão de máquina vetorial, regressão linear múltipla, regressão de potência e regressão robusta) para modelar a relação entre fatores climáticos (precipitação, umidade relativa (mínima e máxima), temperatura (mínima e máxima), velocidade do vento (manhã e noite), evaporação e horas de sol) e a produção de arroz. A significância dos fatores climáticos sobre a produção de arroz foi explorada com o uso do Random Forest (RF). De acordo com os resultados, o estudo identificou que RF é um modelo confiável e preciso para a previsão da produção de arroz no Sri Lanka, determinando que a umidade relativa mínima e a temperatura máxima durante o período de cultivo do arroz são os fatores meteorológicos mais influentes (EKANAYAKE *et al.*, 2021).

Um modelo de previsão espaço-temporal foi proposto por Urrutia *et al.* (2019), para prever a colheita trimestral de cada um dos sete produtores de arroz da região de Luzon, no norte das Filipinas. O estudo demonstrou que o modelo de previsão é superior à previsão ARIMA, mais comumente usada em estudos de séries temporais.

Uma estrutura de rede neural artificial (ANN), que é um algoritmo comum de aprendizado de máquina baseado no modelo do sistema de neurônios humanos, foi determinado por Wickramasinghe et al. (2020) para avaliar as relações entre os componentes climáticos e a colheita

de arroz. Os resultados obtidos a partir da análise revelaram que os rendimentos de arroz previstos em reais têm uma correlação significativa com a precipitação, temperatura máxima e mínima.

Portanto, foi possível observar que, para prever a produção do arroz levando em consideração os fatores climáticos, os modelos lineares generalizados são mais aplicáveis. Diante disso, o presente estudo tem como objetivo determinar um modelo de regressão linear generalizado, considerando variáveis econômicas e meteorológicas, para prever a produção de arroz no estado do Rio Grande do Sul.

Material e Métodos

Nesta pesquisa, foram utilizadas onze variáveis, sendo cinco de origem econômica e seis meteorológicas conforme descrito no Quadro 1, correspondentes a 193 municípios produtores da cultivar no estado. A variável valor da produção de arroz (VP) foi usada como variável resposta e as demais, como variáveis explicativas.

As variáveis econômicas utilizadas foram obtidas no site do Departamento de Economia e Estatística (DEE) em Dados Abertos (DEEDADOS, 2023), considerando os municípios do Rio Grande do Sul que produzem a cultivar. Já as variáveis meteorológicas foram obtidas no site Brazilian Daily Weather Gridded Data (BR-DWGD)(Xavier,2023). Trata-se de uma galeria com dados interpolados para várias variáveis meteorológicas do Brasil, contendo dados de janeiro de 1961 a julho de 2020. Os dados utilizados nesse trabalho são referentes ao ano de 2019.

Chart 1: Definition of variables used in the study.

Variáveis	Definição	Fonte
VP	Valor da produção dada em mil reais (R$).	
AC	Área colhida em hectar (ha).	
AP	Área plantada em hectar (ha).	DEE Dados Abertos
QP	Quantidade produzida em toneladas (ton).	
RM	Rendimento médio da produção (kg/ha).	
RSOL	Média de radiação Solar no ano de estudo (MJ/m2).	
RH	Média da umidade relativa de 2019 (%).	
PR	Média de precipitação no ano de 2019 (mm).	Brazilian Daily
Tmin	Média de temperatura mínima do ano de estudo (°C).	Weather Gridded Data
Tmax	Média de temperatura máximas do ano de 2019 (°C).	
U2	Velocidade do vento a 2 metros (m/s).	

Source: from the authors (2024).

Os dados climáticos passaram por um processo de tratamento antes das análises. Como os dados obtidos abrangiam todo o país, foi necessário selecionar apenas o estado do Rio Grande do Sul. A seleção dos polígonos do estado foi realizada usando o comando geodata::gadm() do pacote *geodata*. O recorte espacial do estado e a aplicação de uma máscara para delimitar as fronteiras foram feitos utilizando os comandos terra::crop() e terra::mask(), respectivamente, ambos do pacote *terra*. Todos estes pacotes citados pertencem ao *software* R. Como resultado, foram obtidas as informações em grade dos municípios produtores de arroz.

A modelagem foi realizada utilizando a metodologia de Modelos Lineares Generalizados (MLG) que é uma extensão dos Modelos Lineares (ML), maiores detalhes sobre esse último modelo podem ser vistos em (Montgomery e Runger, 2009). Os MLG vêm sendo utilizados quando não é possível admitir normalidade para a variável resposta, mas sim qualquer modelo probabilístico que pertença à família exponencial de distribuições (McCullagh e Nelder, 1989).

Esses modelos foram primeiramente apresentados por (McCullagh e Nelder, 1989), contemplando os modelos de regressão linear múltipla, logística, poisson entre outros. Desta forma, o modelo é especificado considerando que a distribuição da variável dependente Y pertence à família exponencial.

O MLG tem a seguinte definição (Cordeiro, 2019):

• Componente aleatório do modelo: a variável resposta é representada por um conjunto de variáveis aleatórias independentes Y_1, ..., Y_n com distribuição pertencente à família exponencial de distribuições de médias μ_1, ..., μ_n ou seja, $E(Y_i) = \mu_i$, i = 1, 2, ..., n;

• Componente sistemático: é constituído pelas variáveis explicativas que entram na forma de uma soma linear de seus efeitos. O componente sistemático ou preditor linear do modelo é dado por $\eta = X\beta$, onde $\eta_i = X_i\beta$ é a componente relativa à i-ésima observação.

• Função de ligação: relaciona o componente aleatório ao componente sistemático, ou seja, é estabelecida pela relação entre o preditor linear η e o valor esperado de um dado y, isto é: $\eta_i = g(\mu_i)$. Sendo g(.) uma função monótona e diferenciável.

Algumas informações descritivas foram apresentadas em tabelas e gráficos, para melhor visualização dos resultados. A normalidade e a homocedasticidade dos resíduos foram testadas pelo teste Shapiro-Wilk e o teste de Breusch-Pagan respectivamente, a independência das variáveis dependentes pelo teste Durbin-Watson. A correlação entre as variáveis foi determinada pelo coeficiente de Spearman. Para ajustar o modelo, foi utilizado o pacote *glm*, utilizando as distribuições gaussiana com função de ligação identidade e Gama com função de ligação identidade e inversa.

Para a seleção das variáveis foi utilizado o método stepwise, e para comparar a qualidade de ajuste dos modelos foi utilizado os critérios Akaike Information Criterion (AIC), deviance, considerando o gráfico de envelope simulado dos resíduos e a distância de Cook.

Todo tratamento dos dados foi feito no *software* R (2022) considerando um nível de significância de 5%.

Resultados e Discussão

O uso de modelos MLG pressupõe independência entre as variáveis aleatórias $Y_1,\ldots,Y_n$. Para verificar essa condição, foi aplicado o teste de Durbin-Watson, cujo resultado indicou independência entre as variáveis (p-valor=0,8725; DW=2,163).

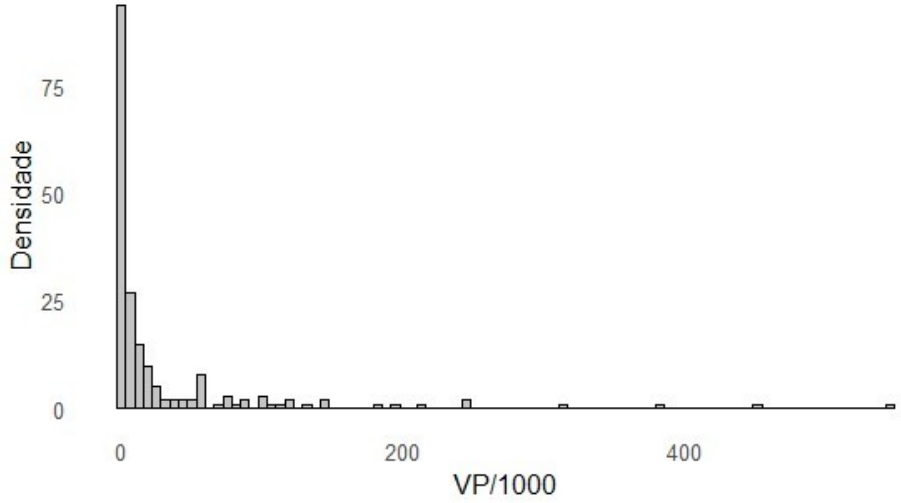
Na Tabela 1 são apresentadas as medidas descritivas das variáveis econômicas. Nota-se que em média os municípios gaúchos obtiveram um valor de produção de R$ 30.759.000, 00 de julho a dezembro de 2019. Observa-se também que a variável resposta apresenta assimetria à direita, o mesmo comportamento acontece com as demais variáveis, com exceção da RM que apresenta uma assimetria à esquerda e um coeficiente de variação consideravelmente menor. O histograma da variável resposta (VP) dado na Figura 1 complementa e corrobora o exposto na Tabela 1. Todas as variáveis apresentam um coeficiente percentílico de curtose caracterizando uma curva platicúrtica.

Table 1: Measures of dispersion, central tendency, skewness, and kurtosis of economic variables.

	Variáveis				
Estatísticas	VP(mil)	AC(ha)	AP(ha)	QP(ton)	RM(kg/ha)
Mínimo	1	1	1	1	1000
1º Quartil	12	5	5	12,2	3000
Mediana	3470	600	600	4209	6800
Média	30759	4967	5049	36674	5571
3º Quartil	20665	3753	3753	27921	7395
Máximo	547699	73308	76319	622385	9100
Desvio Padrão	73330	11251,82	11525,64	86364,54	2428,86
Coef. de variação	2,37	2,27	2,28	2,34	0,44
Assimetria	4,18	3,76	3,85	4,07	-0,62
Curtose	23,74	19,10	20,03	22,50	1,81

Source: from the authors (2024).

Figure 1: Histogram of the response variable.



Source: from the authors (2024).

Verificou-se que em 45 municípios (22, 8%) o valor da produção foi de R$8.000, 00 ou menos e que os 10 munícipios com os maiores valores tiveram um montante mínimo de R$147.074.000,00. Conforme o Atlas Socioeconômico do Rio Grande do Sul (2022), no período 2018 a 2020, dez municípios gaúchos apresentaram produção com média superior a 200.000 toneladas/ano e que os 7 principais municípios juntos são responsáveis por 46% da produção gaúcha.

A umidade relativa apresenta assimetria à esquerda, as demais variáveis apresentam assimetria positiva, assim sendo assimétrica à direita (Tabela 2). Nota-se que a variável RSOL, que já se mostrou importante em outros estudos de arroz como o de Silvio Steinmetz (2013), apresenta um intervalo de variação pequeno, e no ano de estudo a média máxima de radiação solar foi de 16,99. Nota-se também uma caracterização da curva platicúrtica para esses dados.

Table 2: Measures of dispersion, central tendency, skewness, and kurtosis of meteorological variables.

Estatísticas	Variáveis					
	RSOL (MJ/m^2)	RH (%)	PR (mm)	Tmin (ºC)	Tmax (ºC)	U2 (m/s)
Mínimo	15,29	72,10	3,718	12,79	22,09	1,10
1º Quartil	15,76	75,62	4,470	14,48	24,56	1,53
Mediana	15,94	77,28	4,701	15,30	25,53	1,80
Média	16,10	76,97	4,798	15,36	25,50	1,84
3º Quartil	16,49	78,29	5,102	16,21	26,35	2,05
Máximo	16,99	82,55	5,907	20,64	32,94	2,92
Desvio Padrão	0,42	2,28	0,50	1,22	1,28	0,35
Coef. de variação	0,03	0,03	0,11	0,08	0,05	0,19
Assimetria	0,35	-0,19	0,44	0,56	0,81	0,50
Curtose	1,84	2,85	2,59	4,72	7,63	2,71

Source: from the authors (2024).

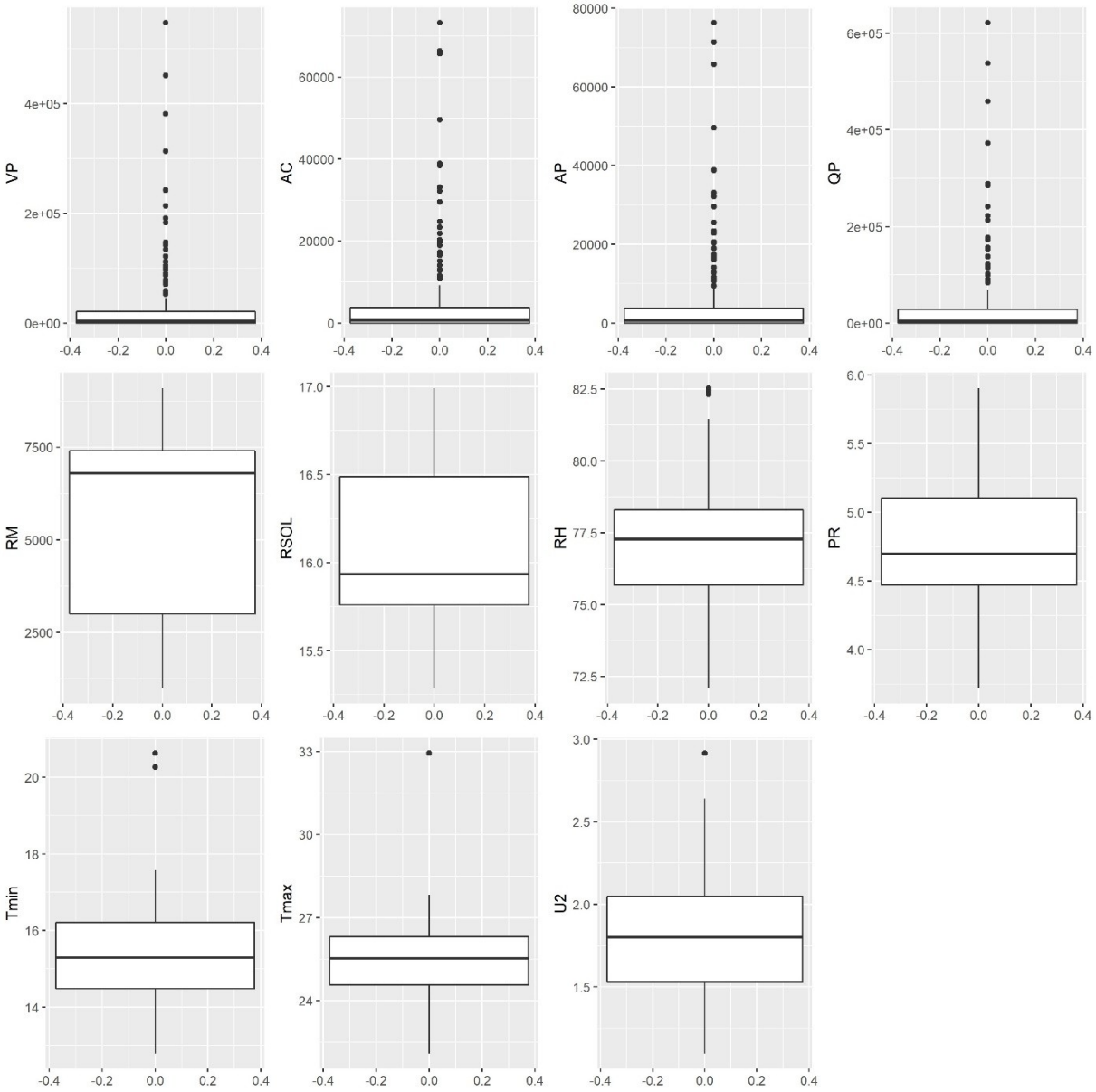
Na Figura 2, são apresentados os boxplots para cada variável. Percebe-se que praticamente todas as variáveis apresentam outliers, exceto as variáveis RM, RSOL e PR. As variáveis que exibem mais pontos discrepantes são VP, AC, AP e QP. Essa discrepância é bastante comum em variáveis econômicas. Outros estudos que também consideraram variáveis econômicas como explicativas, como o de Ribeiro (2019), também apresentaram uma variação expressiva nos dados.

O valor do coeficiente de correlação de Spearman entre as variáveis analisadas é apresentado na Tabela 3. Os valores acima da diagonal correspondem aos coeficientes de correlação, indicando a intensidade e a direção da relação linear entre elas. Observa-se que as covariáveis econômicas AC, AP, QP apresentam uma correlação forte e positiva com a variável de interesse (p-valor < 0,05). Nota-se também uma relação mais fraca entre a variável resposta e a covariável RM, em comparação com as demais covariáveis quantitativas, como mostrado na Figura 3(a), que exibe o diagrama de dispersão das variáveis duas a duas (variável resposta e covariáveis econômicas).

Vale ressaltar que as variáveis AC, AP e QP apresentam uma colinearidade de correlação 1,00, portanto, não há a necessidade mantê-las no modelo, evitando a multicolinearidade nos dados.

Percebe-se, na Tabela 3, que as variáveis meteorológicas têm correlações fracas, algumas negativas, em relação à variável VP. Apenas as variáveis umidade relativa (RH) e a velocidade do vento (U2) são significativas, como mostrado Figura 3(b).

Figure 2: Boxplots of the response variable and quantitative covariates.



Source: from the authors (2024).

Os modelos foram ajustados, considerando VP como variável resposta e, por motivo de multicolinearidade, as variáveis AC e AP foram retiradas da modelagem e, assim, o ajuste foi constituído por uma variável resposta e sete explicativas. Primeiramente, ajustou-se um modelo de regressão múltipla clássico, assumindo normalidade dos resíduos, utilizou-se o método stepwise para seleção das variáveis, obtendo o modelo representado na Tabela 4.

Table 3: Spearman Correlation Coefficient between the analyzed variables.

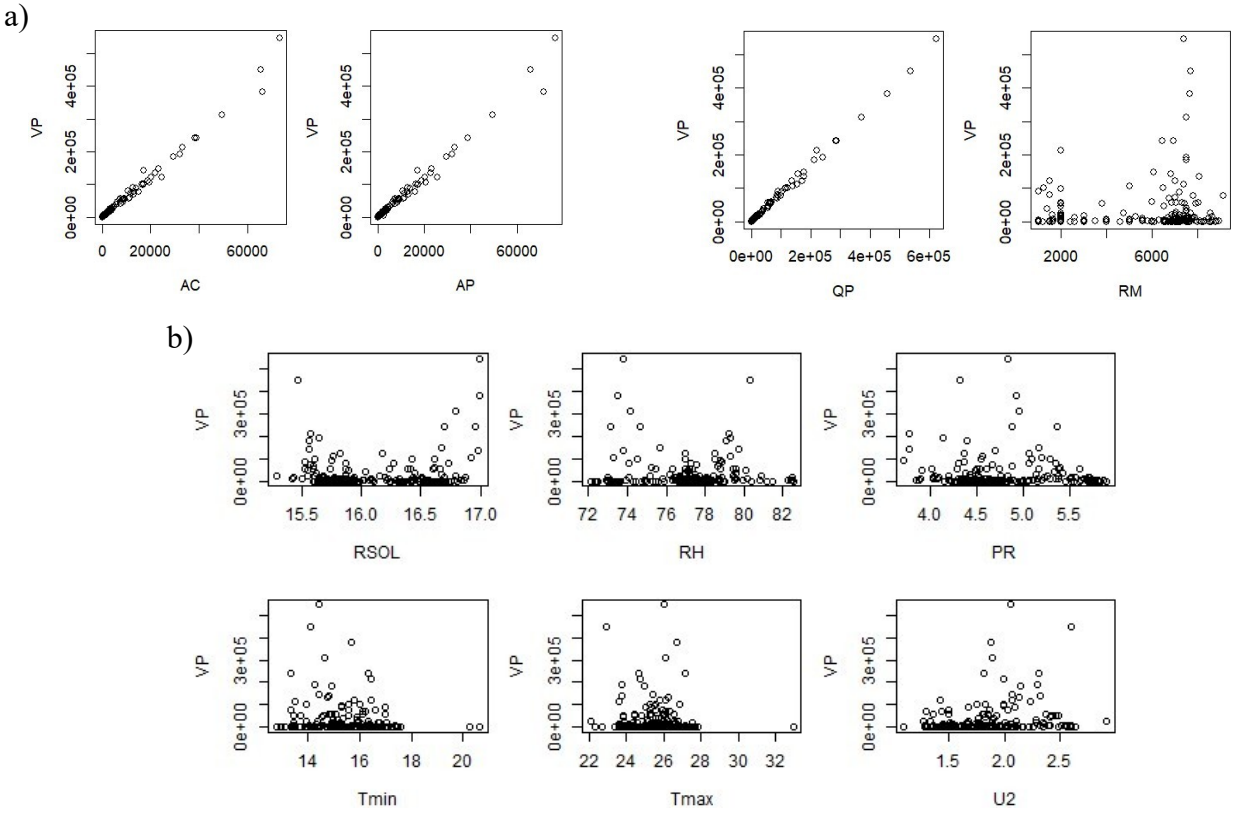
	VP	AC	AP	QP	RM	RSOL	RH	PR	Tmin	Tmax	U2
VP		0,99**	0,99**	1,00**	0,10	-0,08	0,15*	-0,04	-0,04	-0,09	0,27**
AC			1,00**	1,00**	0,10	-0,09	0,16*	-0,04	-0,04	-0,09	0,26**
AP				1,00**	0,10	-0,09	0,15*	-0,04	-0,04	-0,09	0,25**
QP					0,10	-0,09	0,16	-0,04	-0,05	-0,10	0,27**
RM						-0,04	0,05	-0,14	0,00	-0,03	0,10
RSOL							-0,77**	0,63**	0,07	0,33	0,08
RH								-0,62	-0,14	-0,66	0,16*
PR									-0,07	0,34**	0,02
Tmin										0,61	-0,56
Tmax											-0,56
U2											

Source: from the authors (2024).
Legend: **: significance < 0,01, *: significance < 0,05.

Na Tabela 4, observa-se que apenas três covariáveis permaneceram no modelo. Nota-se que teve duas estimativas positivas, sendo elas U2 e RSOL, indicando que quanto maior a radiação solar e mais forte o vento a dois metros de altura, maior será o resultado da variável de interesse (VP). A variável PR influencia negativamente no modelo, tem-se que quanto maior a precipitação de chuva, menor será o valor da produção de arroz. Para os testes de Shapiro-Wilk e Breusch-Pagan, foram obtidos p-valores inferiores ao nível de significância, indicando não normalidade e heterocedasticidade dos resíduos respectivamente.

Observa-se que na Figura 4(a), o gráfico dos resíduos versus os valores ajustados, tem-se vários pontos que não apresentam o mesmo comportamento, o que indica que ocorrem alguns pontos suspeitos a outliers. Na Figura 4(b) é apresentado um gráfico de envelope simulado considerando os resíduos deviance, observa-se que muitos pontos ficam fora das bandas limitantes do gráfico, indicando que não há normalidade dos resíduos assim como foi afirmado com o teste de Shapiro-wilk, tornando inviável o uso da regressão linear. Dessa forma, recorre-se ao método da regressão linear generalizado como forma de contornar as ausências dos pressupostos verificados.

Figure 3: (a) Scatter plot of the response variable with economic covariates (b) Scatter plot of the response variable with meteorological covariates.



Source: from the authors (2024).

Os modelos foram ajustados, considerando VP como variável resposta e, por motivo de multicolinearidade, as variáveis AC e AP foram retiradas da modelagem e, assim, o ajuste foi constituído por uma variável resposta e sete explicativas. Primeiramente, ajustou-se um modelo de regressão múltipla clássico, assumindo normalidade dos resíduos, utilizou-se o método stepwise para seleção das variáveis, obtendo o modelo representado na Tabela 4.

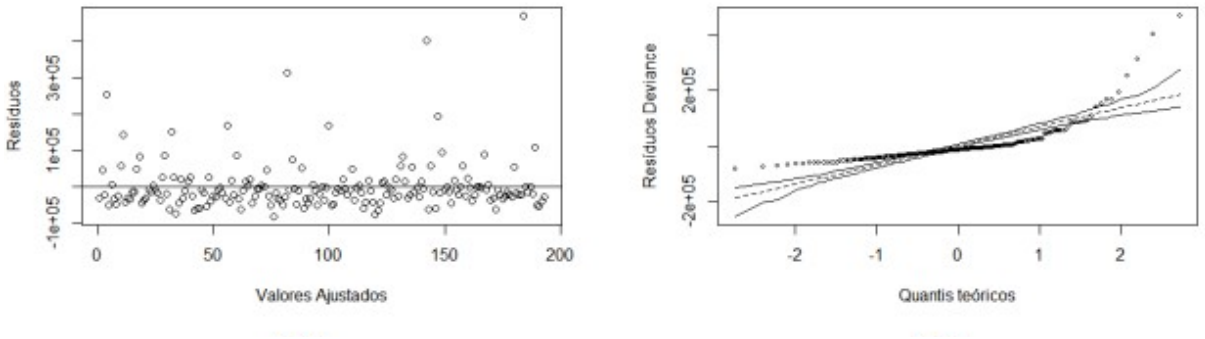
Na Tabela 4, observa-se que apenas três covariáveis permaneceram no modelo. Nota-se que teve duas estimativas positivas, sendo elas U2 e RSOL, indicando que quanto maior a radiação solar e mais forte o vento a dois metros de altura, maior será o resultado da variável de interesse (VP). A variável PR influencia negativamente no modelo, tem-se que quanto maior a precipitação de chuva, menor será o valor da produção de arroz. Para os testes de Shapiro-Wilk e Breusch-Pagan, foram obtidos p-valores inferiores ao nível de significância, indicando não normalidade e heterocedasticidade dos resíduos respectivamente.

Table 4: Estimates of the Effects of the Fitted Gaussian Model.

Covariável	Estimativa	Erro padrão	P-valor
Intercepto	-614047	209537	0,00380 **
U2	41944	14498	0,00426 **
RSOL	44565	14744	0,00286 **
PR	-31162	12373	0,01261 *

Source: from the authors (2024).
Legend: **: significance < 0,01, *: significance < 0,05.

Figure 4: (a) Plot of residuals versus fitted values (b) Simulated envelope plot of deviance residuals for the Gaussian Model.



(a) (b)

Source: from the authors (2024).

Para os MLG, dentre as distribuições testadas, a melhor candidata foi a distribuição Gama com função de ligação canônica identidade e inversa. Na Tabela 5 e na Figura 5 tem-se o ajuste do modelo de regressão generalizado utilizando a distribuição Gama com função de ligação identidade. Assim como no modelo gaussiano, a variável média anual de precipitação se mostrou significativa para o modelo, interferindo negativamente em ambos modelos.

Table 5: Estimates of the Effects of the Gamma Model with Identity Link Function.

Covariável	Estimativa	Erro padrão	P-valor
Intercepto	760874,5	150798,8	< 0, 000
U2	57357,5	8177,9	< 0, 000
PR	-10613,9	1508,6	< 0, 000
Tmin	4995,4	709,5	< 0, 000
RSOL	-20799,3	3783,7	< 0, 000
RH	-6820,2	1304,9	< 0, 000

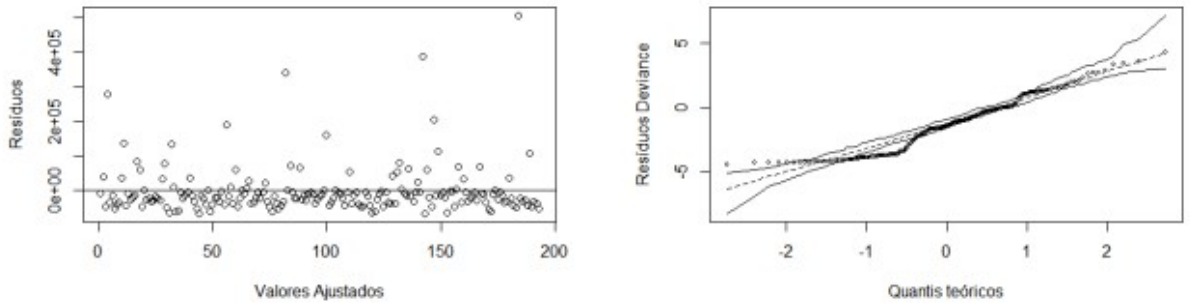
Source: from the authors (2024).

Percebe-se na Tabela 6 que quanto maior a precipitação, radiação solar e a umidade relativa, menor será a média do valor de produção. Porém quanto maior o U2 e Tmin, maior será a média da variável VP. Todas as variáveis foram significativas. Estudos como o de Silvio Steinmetz (2013) apresentaram a radiação solar e temperatura média mínima como variáveis significativas para o desenvolvimento do grão do arroz, logo afetando diretamente no valor da produção da cultivar. O

modelo ajustado considerando a função de ligação identidade para a distribuição Gama é dado pela Equação 1 (AIC = 3693,9):

$$VP = 760874,5 + 57357,5 * U2 - 10613,9 * PR + 4995,4 * Tmin - 20799,3 * RSOL - 6820,2 * RH \quad (1)$$

Figure 5: (a) Plot of residuals versus fitted values (b) Simulated envelope plot of deviance residuals for the Gamma Model with Identity Link Function.



(a) (b)

Source: from the authors (2024).

Na Tabela 6 tem-se o ajuste do modelo usando a distribuição Gama com função de ligação inversa. As variáveis RSOL e PR foram identificadas como importantes para o desempenho da cultivar, pois apresentaram significância no modelo. Isso concorda com resultados de outros estudos, como o apresentado por Nugrahani *et al.* (2021), que também destacaram a precipitação como uma variável que tem uma influência direta no desempenho da cultivar. Assim, o modelo ajustado com base na função de ligação inversa para a distribuição Gama é apresentado na Equação 2 (AIC = 3705,4):

$$\frac{1}{VP} = 4,839e(-4) - 3,782e(-5) * RSOL + 3,448e-5 * PR \quad (2)$$

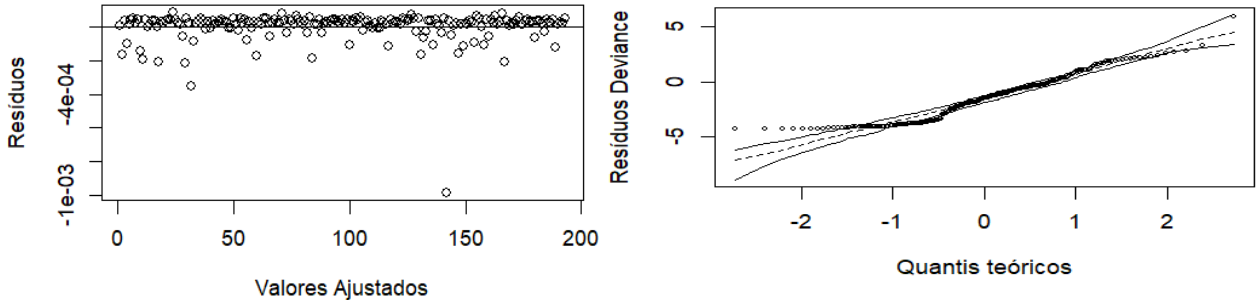
Table 6: Estimates of the Effects of the Gamma Model with Inverse Link Function.

Covariável	Estimativa	Erro padrão	P-valor
Intercepto	4,839e-04	1,389e-04	0,000612 ***
RSOL	-3,782e-05	9,764e-06	0,000148 ***
PR	3,448e-05	1,343e-05	0,010989 *

Source: from the authors (2024).
Legend: ***: significance 0, * significance 0,05.

A Figura 6(a) apresenta o gráfico dos valores ajustados versus os resíduos, onde é observado há vários pontos discrepantes, assim tendo alguns outliers que podem interferir no modelo. Na Figura 6(b) é apresentado o gráfico de envelope considerando os resíduos deviance, em que se observa que nem todos os pontos estão respeitando as bandas limitantes, mostrando um desempenho muito superior comparado ao modelo linear múltiplo.

Figure 6: (a) Plot of residuals versus fitted values (b) Simulated envelope plot of deviance residuals for the Gamma Model with Inverse Link Function.

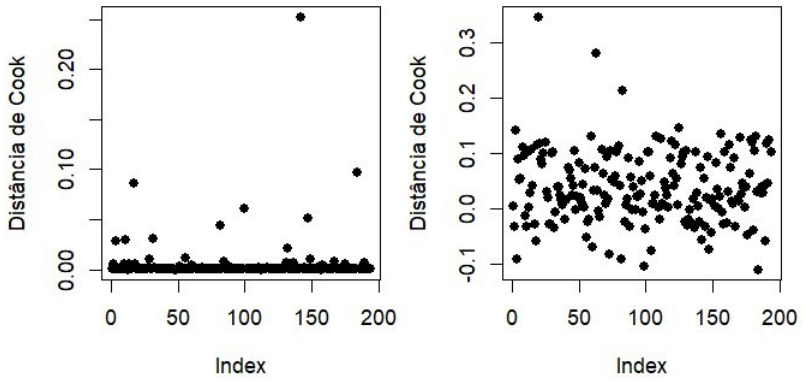


(a) (b)

Source: from the authors (2024).

A análise dos pontos influentes em ambos os modelos foi realizada por meio da distância de Cook, conforme apresentado na Figura 7. O modelo Gama com função de identidade, Figura 7 (a), demonstrou melhor desempenho em comparação com o modelo Gama com função de ligação inversa, uma vez que não apresentou nenhum ponto com valor superior a 0,25.

Figure 7: (a) Cook's distance for the Gamma model with Identity Link Function and (b) Cook's distance for the Gamma model with Inverse Link Function.



(a) (b)

Source: from the authors (2024).

Ademais, foram ajustados os modelos com distribuição Gama e função de ligação logarítmica (Equação 3, AIC = 3889,9), e com distribuição Gaussiana e função de ligação inversa (Equação 4, AIC = 4720,3). No entanto, ambos apresentaram altos valores de AIC, razão pela qual não foram escolhidos para a análise final.

$$logVP = -1{,}6e(-4) + 5{,}69e(-1) * PR + 2{,}78e(-1) * RH + 2{,}8e(-5) * QP \quad (3)$$

$$\frac{1}{VP * = 1{,}03e(-4) - 3{,}58e(-6) * Tmin - 9{,}38e(-6) * U2 - 5{,}065e(-11) * QP} \quad (4)$$

Entre os modelos propostos, o modelo com distribuição Gama e função de ligação identidade apresentou o menor AIC, indicando uma leve melhora no ajuste do modelo. Além disso, a análise de resíduos (Figura 6) e pontos influentes (Figura 7) reforçam que este modelo proporciona o melhor ajuste ao conjunto de dados.

Conclusão

O conjunto de dados referente à produção de arroz no Rio Grande do Sul apresenta valores atípicos que contribuíram para a violação dos pressupostos de uma regressão, mas especificamente a suposição de normalidade.

Portanto, utilizou-se a técnica dos modelos lineares generalizados para suprir tais adversidades, entre as funções da família exponencial, a que obteve um melhor ajuste foi a distribuição Gama, com função de ligação identidade. As variáveis média anual da velocidade do vento, média anual de precipitação, média anual de temperatura mínima, média anual de radiação solar e média anual de umidade relativa foram significativas na modelagem, sendo que a velocidade média do vento e a temperatura mínima, apresentaram sinal positivo, assim afirmando que quanto mais forte o vento e quanto maior a temperatura mínima, maior será o valor da produção, as demais variáveis influenciam negativamente no valor da produção de arroz. O modelo com função de ligação inversa também apresentou as variáveis média anual de radiação solar e média anual de precipitação como significativas.

No entanto, a qualidade do ajuste do MLG não demonstrou ser satisfatória quando analisando os resíduos, o que evidencia uma necessidade de estudos complementares para diagnosticar essa deficiência encontrada, outros trabalhos corroboram com os resultados encontrados neste trabalho. Como sugestão para futuras pesquisas, podem ser utilizadas outras técnicas para melhor auxiliar na compreensão dos fatores que afetam o valor da produção de arroz.

Referências

ATLAS. Arroz: O Rio Grande do Sul é o maior produtor de arroz em casca do Brasil. 7ª ed. 2022. Acesso em: 04 de out. 2022. Disponível em: https://atlassocioeconomico.rs.gov.br/arroz.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos Lineares Generalizados. Minicurso para o 12º SEAGRO e a 52ª Reunião Anual da RBRAS UFSM, Santa Maria, RS,2012. Disponível em: https://docs.ufpr.br/~lucambio/%20CE225/2S2009/livroSeagro.pdf.

DA SILVA, E. V.; JACOBI, L. F. Regressão linear e não linear aplicado ao estudo de casos de AIDS na Região Norte do Brasil. *Ciência E Natura*, v. 42, e27, 2020. https://doi.org/10.5902/2179460X40535

DEEDADOS (2023). Departamento de economia e estatística dados abertos. Disponível em: https://dados.fee.tche.br/

EKANAYAKE, E., WICKRAMASINGHE, L., WELIWATTA, R. Use of regression techniques for rice yield estimation in the north-western province of sri lanka. *Ceylon journal of science*, v. 50, n. 4, p. 439–447, 2021.

https://doi.org/10.4038/cjs.v50i4.7942
MATEUS, W. S.; MATEUS, A. L. S. S.; BUENO FILHO, J. S. DE S. Avaliação por regressão linear múltipla para valor de imóveis residenciais. *Sigmae*, v. 8, n. 2, p. 784–787, 2019. https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/987

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*, 2nd edidion. Chapman and Hall, London, 1989, 526 p.

MONTGOMERY, D. C; RUNGER, G. C. *Estatística aplicada e probabilidade para engenheiros*. 4ª edição. Rio de Janeiro: LTC Editora, 2009, 514 p.

NUGRAHANI, I. D.; SUSANTI, Y.; QONA'AH, N. Modeling of rice production in indonesia using robust regression with the method of moments (mm) estimation. *Nusantara science and technology proceedings*, v. 21, p. 79–87, 2021. http://dx.doi.org/10.11594/nstp.2021.1111

R CORE TEAM.R: A language and environment for statistical computing. R Foundation forStatistical Computing, Vienna, Austria. 2022. Disponível em: https://www.r-project.org/

RIBEIRO *et al.* Soybean production value in the Rio Grande do Sul under the GAMLSS framework. C*ommunications in statistics: case studies, data analysis and applications.* v. 7, n. 2, p. 146-165, 2021. http://dx.doi.org /10.1080/23737484.2020.1852131

SILVA, L. E. M. DA; JACOBI, L. F.; LÚCIO, A. D. Analysis of cases of death due to stroke in the state of Rio Grande do Sul from 1979 to 2014. *Sigmae*, v. 8, n. 2, p. 274–281, 2019. https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/954

SILVIO STEINMETZ, J. B. D. S.; DEIBLER, N. A. Estimativa da produtividade de arroz irrigado em função da radiação solar global e da temperatura mínima do ar. *Ciência Rura*l, Santa Maria, v. 43, n. 2, p. 206-211, 2013. https://doi.org/10.1590/S0103-84782013000200003

URRUTIA, J. D. *et al.* Forecasting rice production in luzon using integrated spatio-temporal forecasting framework. Em: AIP Conference Proceedings, AIP Publishing LLC, v. 2192, p. 09004, 2019. http://dx.doi.org /10.1063/1.5139184

WICKRAMASINGHE, L.; JAYASINGHE, J.; RATHNAYAKE, U. Relationships between climatic factors to the paddy yield in the north-western province of sri lanka. Em: International Research Conference on Smart Computing and Systems Engineering (SCSE), IEEE, p. 223–227, 2020. http://dx.doi.org /10.1109/SCSE49731.2020.9313047

XAVIER, A. C. Brazilian daily weather gridded data (br-dwgd), 2023. Disponível em: https://sites.google.com/site/alexandrecandidoxavierufes/brazilian-daily-weather-gridded-data