Análise de Regressão na Produtividade de Soja

Arthur Hintz

11 jul 2024

- Resumo
- Introdução
- Dados
- Análise Descritiva
- Ajustes dos Dados
- Modelo Ajustado
- Analise de diagnóstico
  - Alavancagem
  - DFFIT
  - Distância de Cook
  - Resíduo
  - Envelope Simulado
- Suposições do modelo
  - Teste RESET
  - Teste t para a média dos errros
  - Teste de Bressch-Pagan
  - Teste de Durbin-Watson
  - Fatores de Inflação de Variância
  - Teste Jarque-Bera
- Predição
- Conclusão
- Trabalhos futuros

Resumo

O trabalho tem como objetivo realizar uma análise de regressão linear múltipla para estimar a produtividade de soja em (Kg/ha) com base nos principais fatores que a influenciam. A análise será dividida em quatro etapas principais: análise descritiva, ajuste do modelo, diagnóstico de influência e teste das suposições do modelo.

Introdução

Segundo a Confederação da Agricultura e Pecuária do Brasil (CNA), o Produto Interno Bruto (PIB) do Agronegócio corresponde a 23,8% em 2023, sendo a soja a commodity de maior valor de produção no Brasil, de acordo com dados divulgados pelo IBGE, link de acesso. Na produção de soja, muitos fatores influenciam o resultado, sendo vários deles incontroláveis, como fatores climáticos. Este ano,

por exemplo, houve um aumento de 70,83% na produtividade em relação ao ano anterior, provavelmente relacionado ao volume de precipitação.

Dessa forma, torna-se necessário estimar a produtividade da soja e verificar quais são as principais variáveis que a influenciam, permitindo realizar predições da safra e estimar o potencial de produção a nível nacional.

Inicialmente, será realizada uma análise descritiva dos dados para compreender as variáveis envolvidas e suas inter-relações. Em seguida, o modelo de regressão linear múltipla será ajustado para identificar os fatores mais significativos que afetam a produtividade da soja. O diagnóstico de influência ajudará a identificar pontos de dados que têm um impacto desproporcional no ajuste do modelo, possibilitando a correção ou análise adicional desses pontos. Finalmente, as suposições do modelo de regressão linear serão testadas para garantir a validade das conclusões obtidas.

Os dados utilizados neste estudo foram disponibilizados pela empresa Crops Team e foram coletados a partir de experimentos com cultivares de soja realizados em diversos locais do estado do Rio Grande do Sul durante as safras de 2021/2022 e 2022/2023.

Este estudo é importante porque permite identificar os principais determinantes da produtividade da soja, fornecendo insights valiosos para a tomada de decisões agrícolas e a otimização dos rendimentos das cultivares. Ao compreender melhor os fatores que influenciam a produtividade, produtores e pesquisadores podem implementar práticas agrícolas mais eficazes.

/ Dados

O banco de dados, inicialmente, continha informações dos 4 blocos de ensaios para cada cultivar. Posteriormente, foi calculada a média dos blocos por cultivar, resultando em 1513 observações e 33 variáveis. Entretanto, após o precesso de filtragens e tratamento de valores faltantes, esses números foram reduzidos. As variáveis do banco de dados incluem informações sobre as cultivares, características do local, dados climáticos durante o período dos experimentos e componentes químicos do solo.

- **Cultivares:** Informações sobre as diferentes variedades de soja utilizadas nos experimentos.
- **Localização:** Características geográficas dos locais onde os experimentos foram conduzidos.
- **Dados Climáticos:** Informações sobre precipitação, temperatura e outras condições climáticas durante o período dos experimentos.
- **Componentes Químicos do Solo:** Dados sobre a composição química do solo, incluindo níveis de nutrientes e pH.

Dentre essas principais características, as variáveis mais significativas utilizadas no modelo foram:

1. `Terras`: divididas em (Altas ou Baixas)
2. `Ambiente`: dividido em (Sequeiro ou Irrigado)
3. `Cultura_Ant`: (“arroz e pousio”,“aveia”, “aveia branca”, “aveia e centeio”, “aveia e ervilhaca”, “azevem”, “cevada”, nabo”)
4. `P_base`: Quantidade de adubação de Fósforo
5. `N_base`: Quantidade de adubação de Nitrogênio
6. `Produtividade`: Produtividade de soja (Kg/ha)
7. `GMR`: Grupo de maturação relativo
8. `Espacamento`: Espaçamento entre linhas do plantio de soja
9. `Temperatura_Max`: média da temperatura máxima durante o período

10. `PH`: PH do solo
11. `M.O.(%)`: Matéria orgânica (%)
12. `Epoca_de_semeadura`: Data de plantio

Sendo as primeiras colunas e observações dadas por:

X	Safra	COD_PROD	Local	Cod_Estacao_Met	Altitude	Terras	Ambiente	Cultivar	Cultura_Ant	Epoca_de_semeadura	N_base
390	2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	AS 3595 I2X	azevem	2021-11-23	13.5
391	2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	AS 3615 I2X	azevem	2021-11-23	13.5
392	2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	BMX COMPACTA IPRO	azevem	2021-11-23	13.5
393	2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	BMX CROMO TF IPRO	azevem	2021-11-23	13.5
394	2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	BMX LOTUS IPRO	azevem	2021-11-23	13.5

Análise Descritiva

As análises dos dados referem-se a um processo crítico em relação produtividade de soja no RS. Dessa forma, foi verificado medidas de tendência central, medidas de disperção, as relações entre as variáveis e suas distribuições.

Data summary

Name	dados
Number of rows	1513
Number of columns	33
_____	
Column type frequency:	
character	10
numeric	22
POSIXct	1
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	min	max	empty	n_unique	whitespace
X	0	3	4	0	1513	0
Safra	0	9	9	0	2	0
COD_PROD	0	5	5	0	34	0
Local	0	5	23	0	27	0
Cod_Estacao_Met	0	4	4	0	18	0
Terras	0	5	6	0	2	0
Ambiente	0	8	8	0	3	0
Cultivar	0	5	25	0	184	0
Cultura_Ant	0	2	17	0	13	0
PRODUTOR/PARCEIRO	0	3	26	0	32	0

Variable type: numeric

skim_variable	n_missing	mean	sd	p0	p25	p50	p75	p100
Altitude	0	304.69	217.97	3.00	105.00	288.00	489.00	688.0
N_base	191	12.46	8.56	0.00	6.00	9.20	17.20	40.00
P_base	191	74.45	28.20	40.00	56.00	64.40	92.00	135.0
K_base	191	41.47	32.79	0.00	0.00	45.50	60.00	112.5
Produtividade	0	2630.00	1417.06	185.00	1512.70	2416.70	3622.57	6898
GMR	32	5.83	0.43	4.90	5.50	5.80	6.10	8.10
N_Plantas	474	77.28	25.36	12.00	59.75	74.50	92.25	172.0
Espacamento	53	0.47	0.04	0.40	0.45	0.45	0.45	0.58
Area_colhida	247	3.37	0.96	0.90	2.70	3.60	4.05	5.40
Populacao	247	18.12	9.21	0.00	15.25	20.30	24.07	43.98
Temperatura_Max	0	24.42	1.23	22.08	23.59	24.48	25.08	28.10
Precipitacao_Total	0	389.34	88.58	158.70	315.90	372.80	451.20	735.2

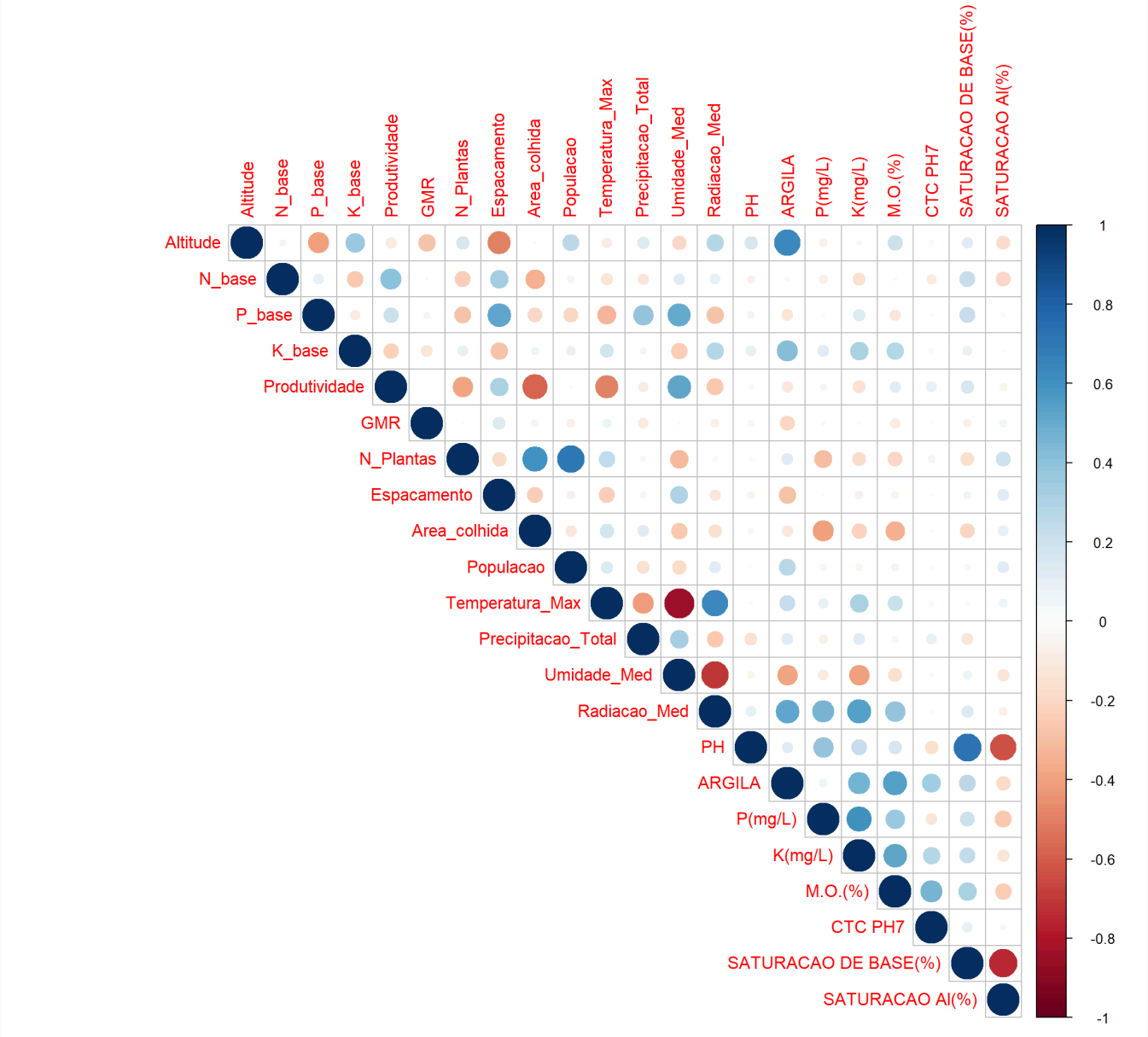
skim_variable	n_missing	mean	sd	p0	p25	p50	p75	p100
Umidade_Med	0	68.61	6.20	52.93	64.45	68.68	72.00	80.19
Radiacao_Med	0	24285.63	2940.83	20411.79	22770.84	23768.85	24520.27	3859
PH	0	5.18	0.35	4.50	5.00	5.10	5.30	6.30
ARGILA	18	43.19	20.31	4.00	27.00	44.00	59.00	85.00
P(mg/L)	0	26.14	41.91	2.50	9.00	19.00	27.00	359.2
K(mg/L)	0	130.42	109.59	28.00	72.00	88.00	180.00	636.0
M.O.(%)	0	2.51	0.83	1.00	1.90	2.50	2.90	5.90
CTC PH7	0	14.09	4.21	6.90	11.40	14.00	15.40	35.50
SATURACAO DE BASE(%)	0	62.70	14.10	27.00	54.00	65.00	76.00	86.00
SATURACAO Al(%)	0	3.38	5.74	0.00	0.00	2.10	4.00	34.00

Variable type: POSIXct

skim_variable	n_missing	min	max	median	n_unique
Epoca_de_semeadura	0	2021-09-16	2022-12-14	2022-11-06	40

Podemos verificar valores faltantes em algumas variáveis no banco de dados, dessa forma, foram retiradas essas observações devido a falta de informação para o preenchimento correto desses NA's. Além disso, podemos ter uma ideia da média e da distribuição dos dados

Com o gráfico a seguir, podemos verificar as principais correlações entre as variáveis numéricas e já procurar possiveis casos de multicolinearidade e variáveis que podem ser mais significativas para estimar a produtividade de soja

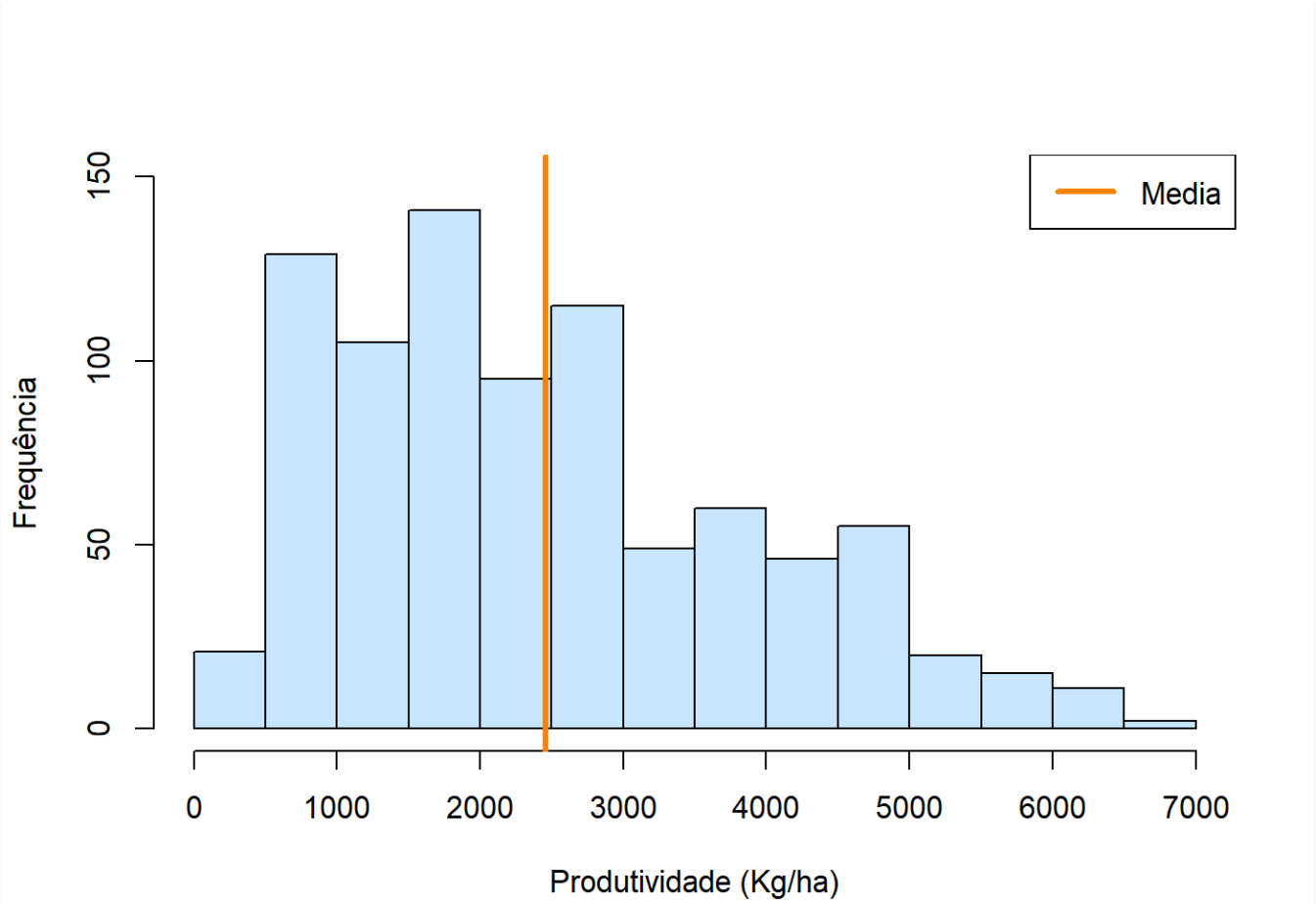


O gráfico de correlação entre as variáveis pode ser interpretado a partir do tamanho dos círculos e da cor. Quanto maior o círculo e mais escura é a cor mais forte é a correlação, também, se puxar mais para o azul é positiva, se for vermelha é negativa.

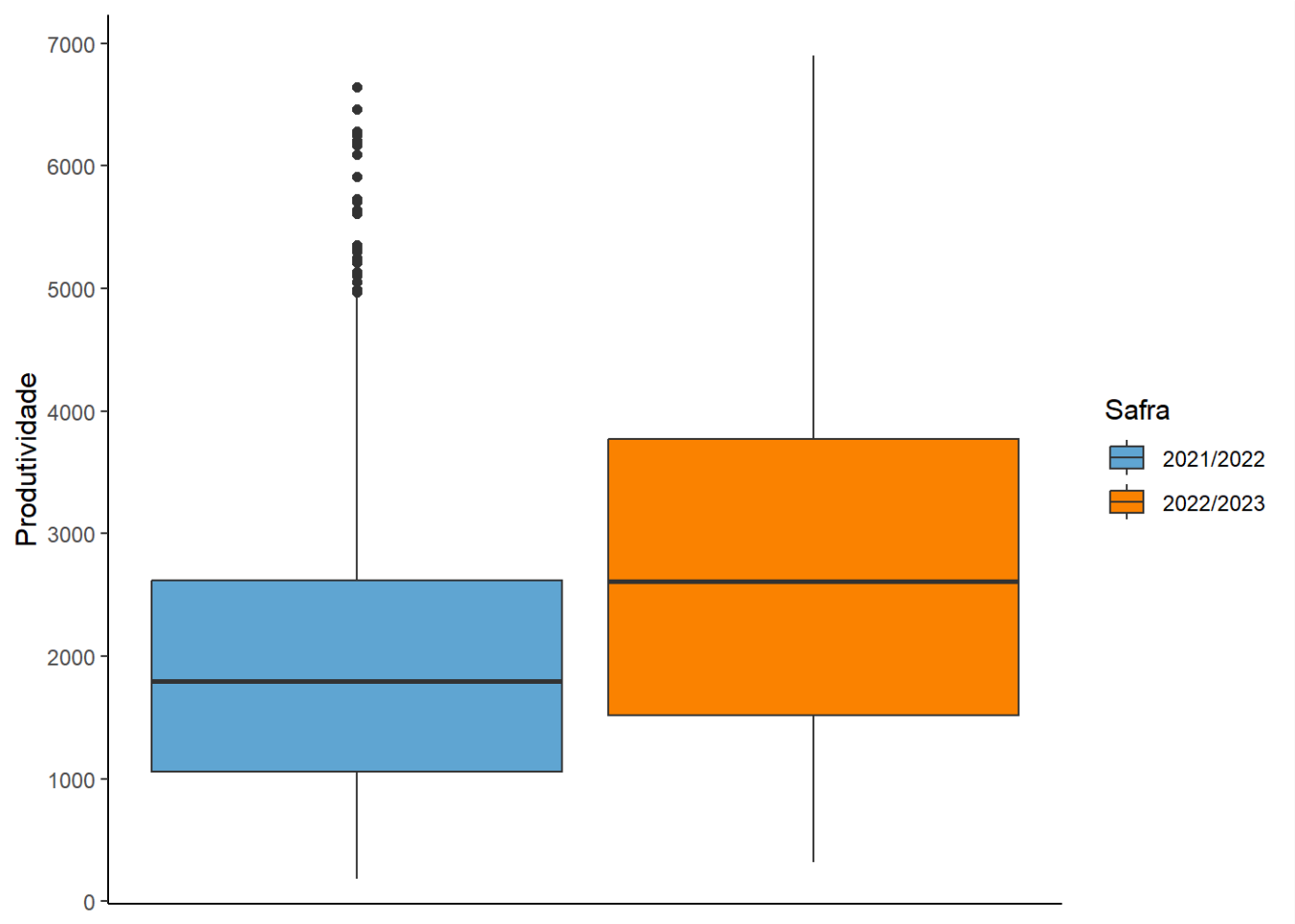
Principais medidas da variável de interesse e um gráfico para mostrar frequência de produtividade

Média da produtividade: 2458

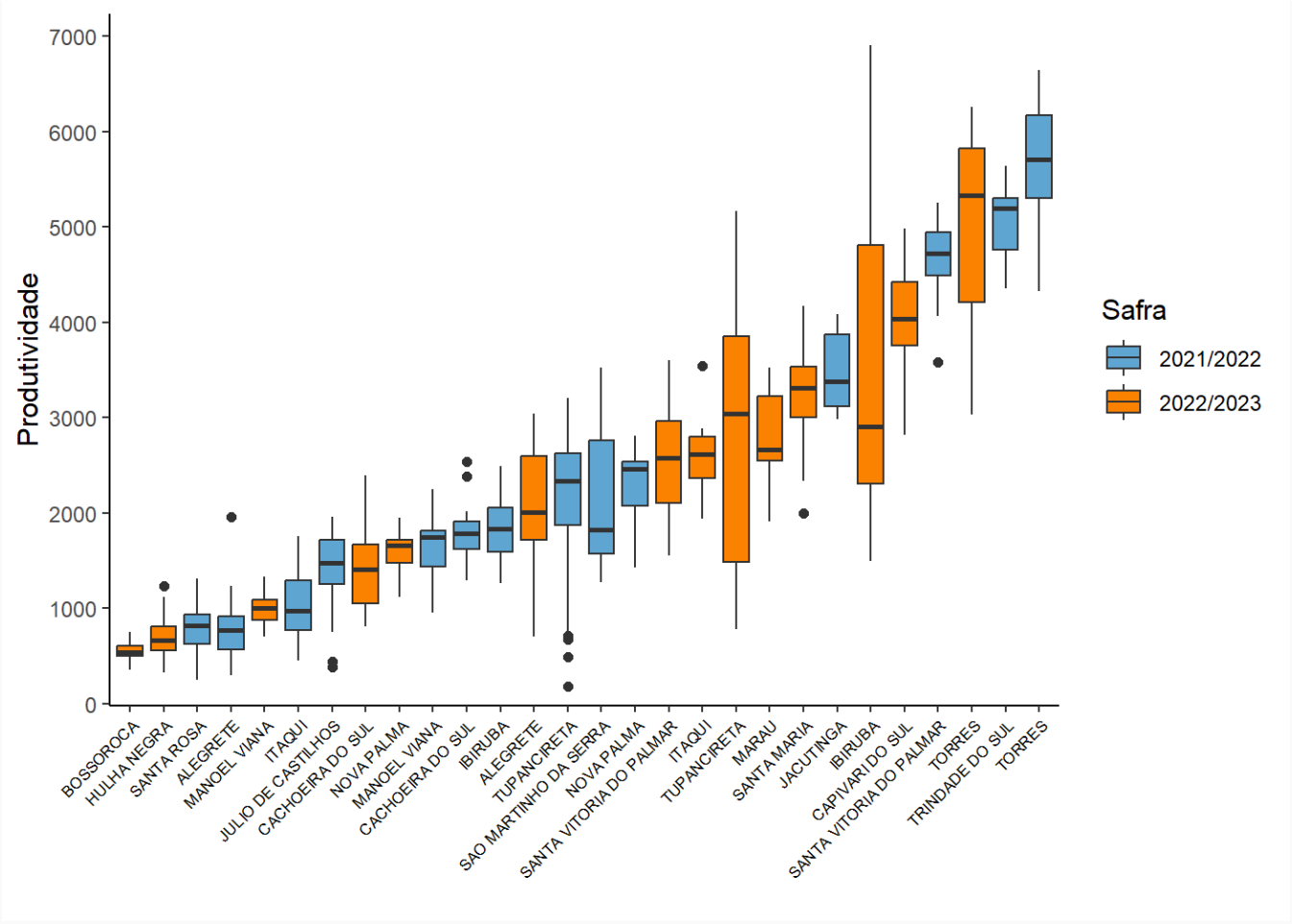
Desvio padrão da produtividade: 1434



O gráfico de boxplot mostra a relação entre os anos das safras com a produtividade



Nota-se um diferença de produtividade entre as duas safras, de certa forma a variável `Safra` deveria ser significativa para a explicação da produtividade de soja, no entanto ja tem relação com outras variáveis do modelo. Além disso, deve explicar a alta variabilidade dos erros, conforme será apresentado nas suposições do modelo.



A partir do gráfico percebe-se a alta variabilidade de produtividade entre os locais de ensaio nas duas safras.

Ajustes dos Dados

1 - Inicialmente foi removida uma cultivar experimental e cultivares com grupos de maturação relativos maiores que 7

2 - Os locais Santa Rosa e Jacutinga apresentaram ser pontos influêntes para o modelo. Logo para o ajuste foi melhor remover essas observações

3 - Locais os quais não tiveram uma cultura antes do plantio de soja ou tiveram mix também foram pontos de alavancagem para o modelo.

4 - Criação da variável `mês`, relacionada a data de plantio, ou seja, invés de ter dia e mês, tem se apenas o mês

5 - Foi alterado a variável `Terras`, em que Baixas recebe 0 e Altas recebe 1. A variável `Ambiente`, “irrigado” = 1 e “sequeiro” = 0

Os dados depois de filtrados e selecionados as variáveis importantes para o modelo, é dado por:

Terras	Ambiente	Cultura_Ant	P_base	N_base	Produtividade	GMR	Espacamento	Temperatura_Max	PH	M.O.(%)	mes
0	1	azevem	135.0	9.0	4330.1250	6.1	0.50	23.26305	4.9	2.0	nov

Terras	Ambiente	Cultura_Ant	P_base	N_base	Produtividade	GMR	Espacamento	Temperatura_Max	PH	M.O.(%)	mes
1	0	aveia	69.0	6.0	2585.2788	5.8	0.45	24.59763	5.0	2.6	nov
1	1	aveia	43.0	17.2	4275.1000	5.4	0.45	24.76108	4.9	3.0	out
0	0	azevem	69.0	13.5	405.2235	5.7	0.45	25.07801	4.8	1.2	nov
1	1	trigo	40.0	16.0	3592.3250	5.8	0.45	25.35516	5.1	2.9	nov
1	0	trigo	40.0	16.0	906.2000	5.4	0.45	25.35516	5.1	2.9	nov
1	1	aveia e ervilhaca	57.5	5.0	2897.3250	6.4	0.45	24.47887	6.3	2.8	out
1	1	aveia e ervilhaca	57.5	5.0	2735.3250	5.9	0.45	24.47887	6.3	2.8	out
1	1	aveia e ervilhaca	57.5	5.0	1058.9250	4.9	0.45	23.69126	4.8	2.7	nov
0	0	azevem	56.0	6.0	3135.2500	5.2	0.45	23.68543	5.3	2.0	dez

Modelo Ajustado

Inicialmente, foi selecionada as variáveis do modelo pelo algoritmo de Stepwise, determinando aquele com menor AIC

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34062.508754	908.227227	37.5043907	0.0000000
Terras1	-1494.382152	81.064383	-18.4345098	0.0000000
Ambiente0	-2410.469323	60.900930	-39.5801725	0.0000000
Cultura_Antaveia	-222.616571	127.881313	-1.7408061	0.0821773
Cultura_Antaveia e centeio	-1633.721856	159.065955	-10.2707198	0.0000000
Cultura_Antaveia e ervilhaca	-3278.926009	145.752264	-22.4965700	0.0000000
Cultura_Antazevem	-1333.979455	99.802047	-13.3662534	0.0000000
Cultura_Antcevada	-314.252257	145.808254	-2.1552433	0.0314989
Cultura_Antnabo	-2098.207963	173.752054	-12.0758743	0.0000000
Cultura_Anttrigo	-74.332602	115.227491	-0.6450943	0.5190872
P_base	5.817805	1.336358	4.3534770	0.0000155
N_base	-36.833253	3.803653	-9.6836523	0.0000000
GMR	177.136001	43.457706	4.0760551	0.0000513
Espacamento	-13410.385353	969.980868	-13.8254122	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
Temperatura_Max	-1080.879735	29.997418	-36.0324257	0.0000000
PH	589.197131	73.340645	8.0337053	0.0000000
`M.O.(%)`	477.670294	33.857294	14.1083424	0.0000000
mesnov	-646.319615	77.359217	-8.3547848	0.0000000
mesout	-486.693245	74.382306	-6.5431319	0.0000000

O Modelo de regresssão linear múltipla, é expressado pela equação:

$$Y = \beta_0 + \sum_{i=1}^{18} \beta_i X_i + \epsilon$$

Sendo:

$$Y = 34062.5 - 1494.4X_1 - 2410.5X_2 - 222.6X_3 - 1633.7X_4 - 3278.9X_5 - 1334X_6 - 314.3X_7 - 2098.2X_8 - 74.3X_9 + 5.8X_{10} - 36.8X_{11} + 177.1X_{12} - 13410.4X_{13} - 1080.9X_{14} + 589.2X_{15} + 477.7X_{16} - 646.3X_{17} - 486.7X_{18} + \epsilon$$

Em que:

X_1 = Terras, $x \in \{0, 1\}$
X_2 = Ambiente, $x \in \{0, 1\}$
X_3 = Aveia, $x \in \{0, 1\}$
X_4 = Aveia e centeio, $x \in \{0, 1\}$
X_5 = Aveia e ervilhaca, $x \in \{0, 1\}$
X_6 = Azevém, $x \in \{0, 1\}$
X_7 = Cevada, $x \in \{0, 1\}$
X_8 = Nabo, $x \in \{0, 1\}$
X_9 = Trigo, $x \in \{0, 1\}$
X_{10} = P_base, $x \in [40, 135]$
X_{11} = N_base, $x \in [5, 40]$
X_{12} = GMR, $x \in [4.9, 6.7]$
X_{13} = Espacamento, $x \in [0.4, 0.575]$
X_{14} = Temperatura_Max, $x \in [22.32, 27.07]$
X_{15} = PH, $x \in [4.8, 6.3]$
X_{16} = M.O.(%), $x \in [1, 4.2]$
X_{17} = Novembro, $x \in \{0, 1\}$
X_{18} = Dezembro, $x \in \{0, 1\}$

Interpretação dos betas:

- β_1, significa que terras altas produzem -1494.4 kg/ha a menos que terras baixas.
- β_2, significa que ambientes sequeiros produzem -2410.5 kg/ha a menos que ambientes irrigados

- $\beta_3, \cdots, \beta_9$, quanto maior o beta maior é significante para o aumento de produtividade, ou seja, neste caso o plantio de trigo antes da soja melhora a protuvidade.
- β_{10}, a cada um 1 kg/ha de fósforo, dentro do intervalo de X_{10} estabelecido, aumenta 5.8 kg/ha na produtividade de soja.
- β_{11} a cada um 1 kg/ha de nitrogênio, dentro do intervalo de X_{11} estabelecido, diminui -36.8 kg/ha na produtividade de soja.
- β_{12}, quanto maior for o GMR, maior será a produtividade
- β_{13}, espaçamentos menores tem mais incremento na produtividade
- β_{14}, temperaturas mais altas diminuem a produtividade.
- β_{15}, o PH do solo tem um fator positivo na produtividade
- β_{16}, a cada um 1% de matéria orgânica, aumenta em 477.7 kg/ha de soja
- β_{17} e β_{18}, o plantio no mês de outubro em relação ao mês de novembro tem um acréscimo de aproximadamente 160 kg/ha.

O coeficiente de determinação, R^2, é dado por 0.9106993, ou seja, significa que 91.1% da variabilidade na produtividade da soja pode ser explicada pelas variáveis incluídas no modelo de regressão.

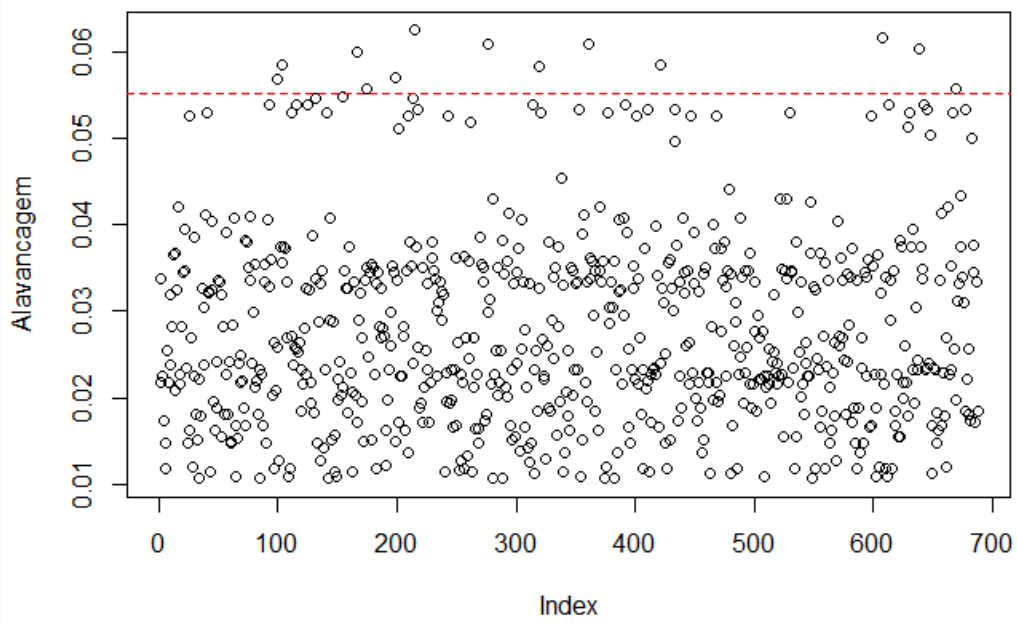
/ Analise de diagnóstico

As principais observações influentes podem ser visualizadas na tabela a seguir

Terras	Ambiente	Cultura_Ant	P_base	N_base	Produtividade	GMR	Espacamento	Temperatura_Max	PH	M.O.(%)	mes
0	1	azevem	135.0	9.0	4330.125	6.1	0.50	23.26305	4.9	2.0	nov
1	1	aveia e ervilhaca	57.5	5.0	1058.925	4.9	0.45	23.69126	4.8	2.7	nov
1	0	aveia e centeio	90.0	6.0	868.800	5.8	0.45	23.95177	4.9	1.4	nov
1	0	azevem	54.6	7.8	1907.925	5.2	0.45	22.68672	5.4	2.5	nov
1	0	aveia e centeio	90.0	6.0	750.800	5.3	0.45	23.95177	4.9	1.4	nov
1	0	aveia e centeio	90.0	6.0	694.150	5.0	0.45	23.95177	4.9	1.4	nov
1	0	aveia e centeio	90.0	6.0	1058.175	6.4	0.45	23.95177	4.9	1.4	nov
1	0	aveia e centeio	90.0	6.0	957.900	6.0	0.45	23.95177	4.9	1.4	nov

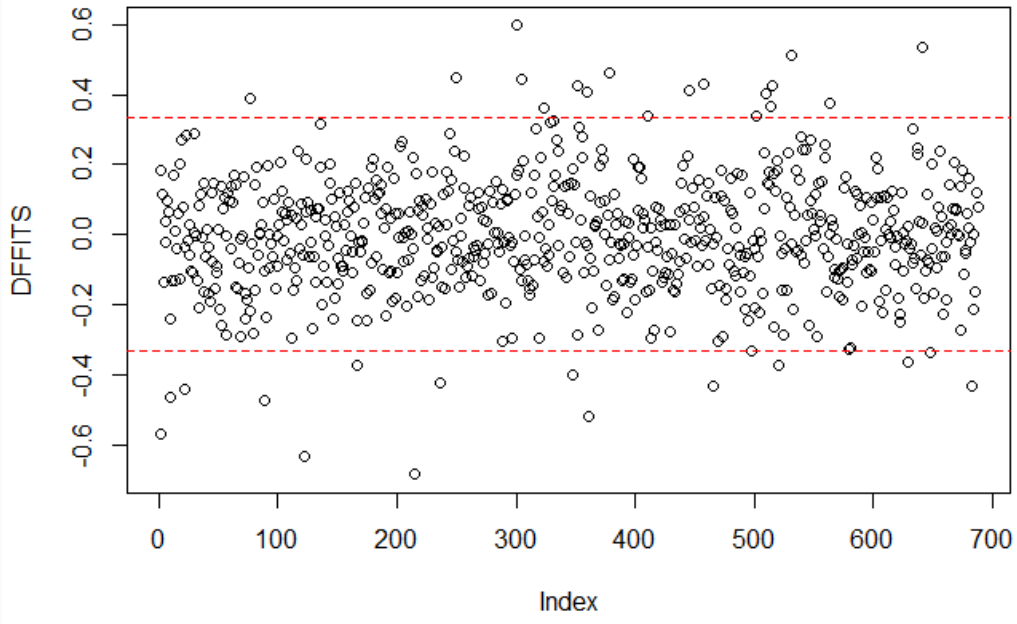
Apesar de ainda possuir pontos de influência não afetam no ajuste do modelo

// Alavancagem



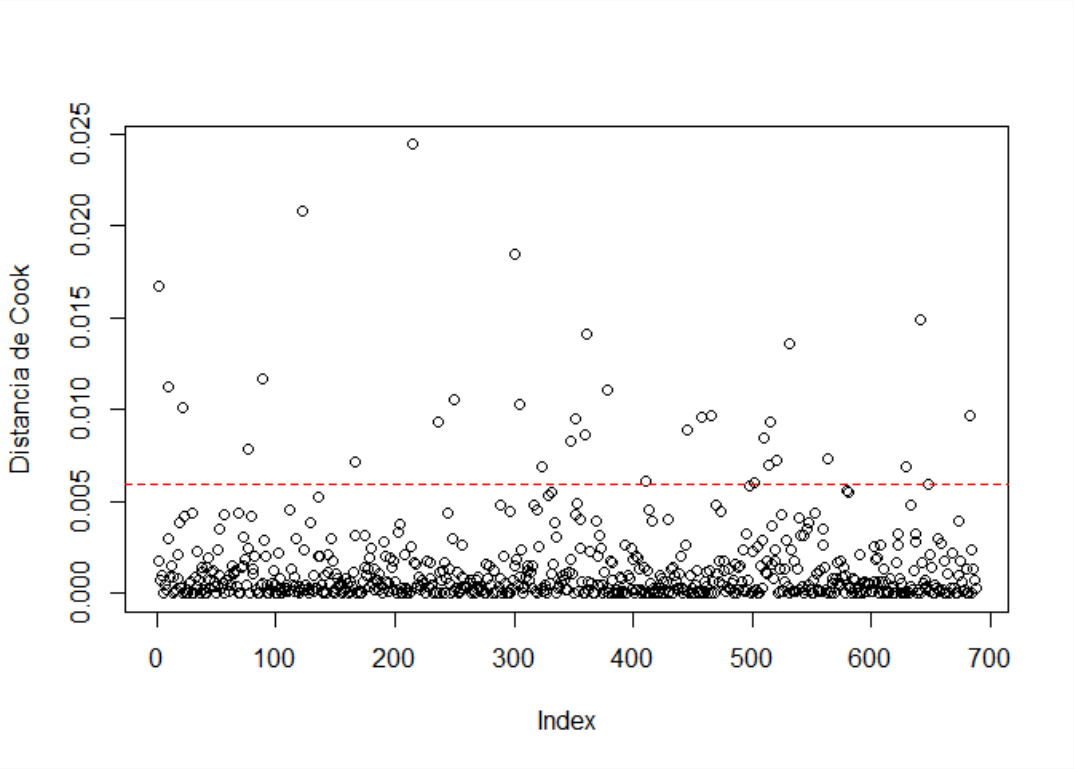
Como de esperado ficou alguns pontos de alavancagem, mas a retirada deles não afetaram o R^2 do modelo

// DFFIT



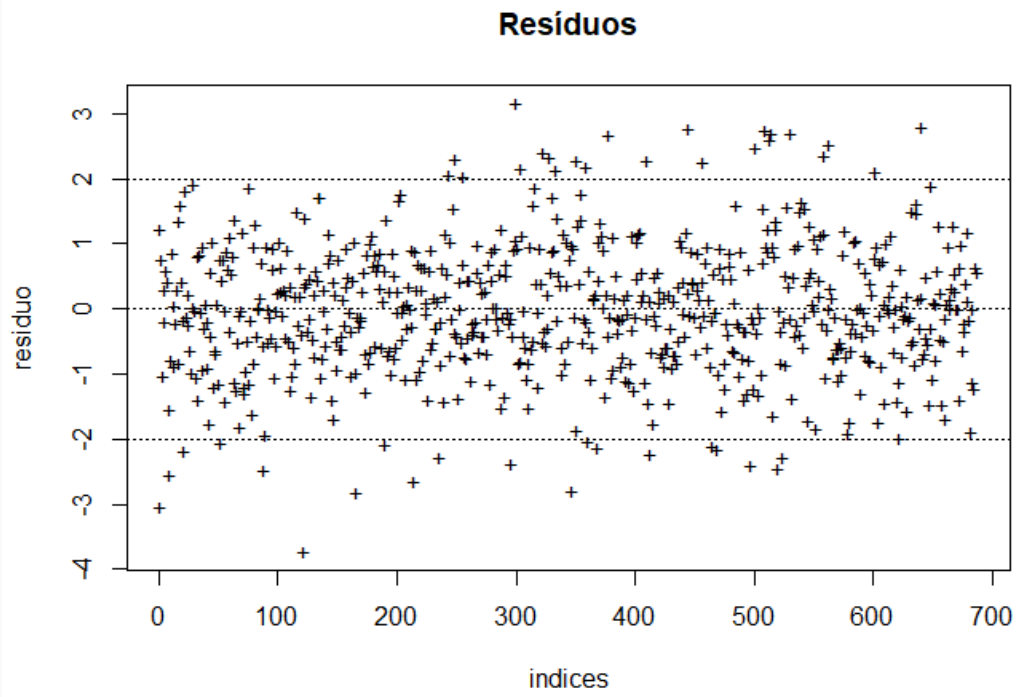
É possível observar que alguns pontos têm valores dos DFFITS acima da linha de referência, mas após análises, eles não apresentaram influência na regressão

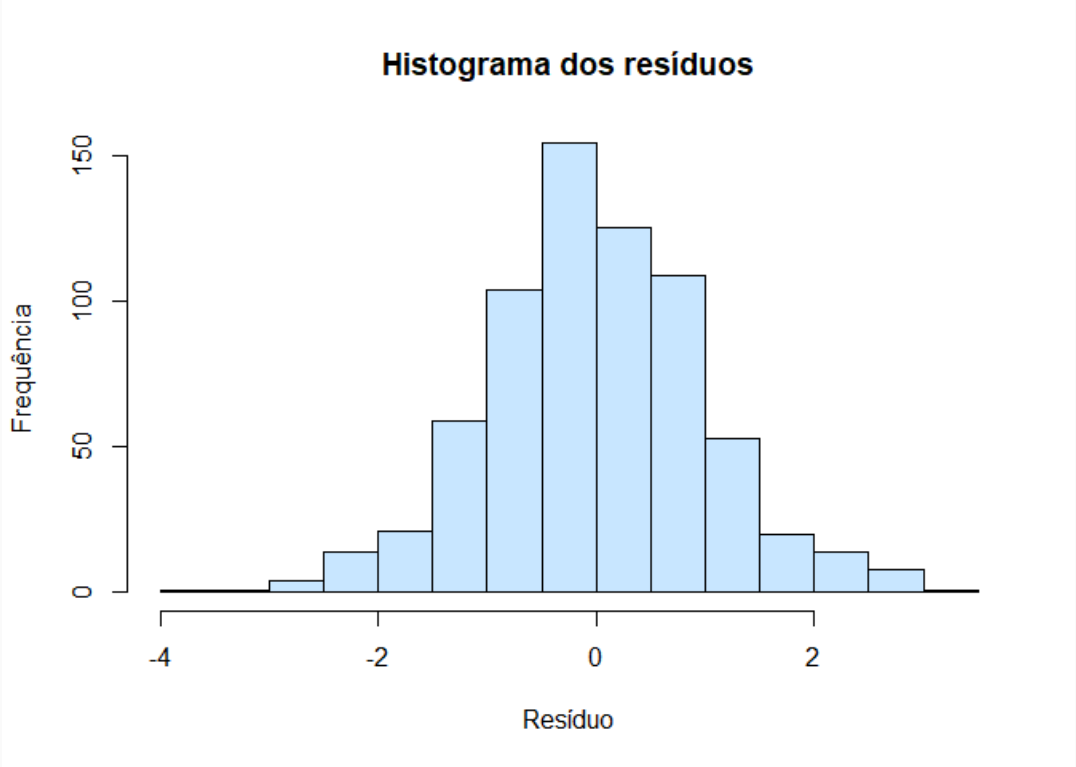
// Distância de Cook



Os mesmos pontos identificados anteriormente aparece no gráfico da distância de cook, mas a remoção dos mesmos não afeta significativamente nos coeficientes da regressão.

// Resíduo

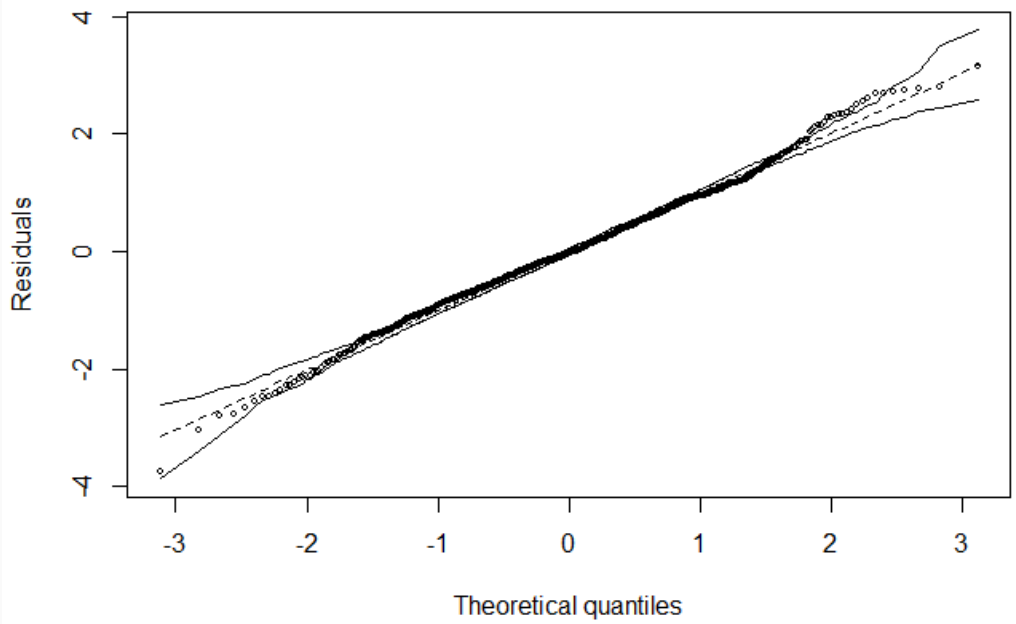




Como esperado, a distribuição dos resíduos tende a se aproximar de uma distribuição normal com média 0. Esta suposição será verificada a seguir através de testes de normalidade.

// Envelope Simulado

Baseado nos resíduos studentizados para verificação de normalidade



A partir do gráfico podemos ver alguns pontos fora do intervalo, mas como n = 688, é esperado a 5% que até 34.4 pontos fiquem fora do intervale e após testes os resíduos seguem normalidade conforme é esperado.

/ Suposições do modelo

- [S0] O modelo esta corretamente específicado
- [S1] A média dos erros é zero
- [S2] Homoscedasticidade dos erros
- [S3] Não autocorrelacão
- [S4] Ausência de Multicolinearidade
- [S5] Normalidade dos erros

// Teste RESET

$$\begin{cases} H_0 : \text{O modelo esta corretamente especificado} \\ H_1 : \text{O modelo não esta corretamente especificado.} \end{cases}$$

```
##
##  RESET test
##
## data:  fit
## RESET = 3.8991, df1 = 2, df2 = 667, p-value = 0.02072
```

Conforme o tese RESET, utilizado para verificar se o modelo está corretamente específicado, não rejeita-se H_0 devido p-valor = 0.0207244 > α = 0.01, ou seja, O modelo esta corretamente especificado.

// Teste t para a média dos errros

$$\begin{cases} H_0: \text{ A média dos erros é igual a zero} \\ H_1: \text{média dos erros é diferente de zero.} \end{cases}$$

```
##
##  One Sample t-test
##
## data:  resid(fit)
## t = -1.9068e-15, df = 687, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -31.82613  31.82613
## sample estimates:
##     mean of x
## -3.090807e-14
```

Conforme o teste T de Student, não rejeita-se H_0 devido ao p-valor = 1 > α = 0.01. Dessa forma, a média dos erros é igual a zero.

// Teste de Bressch-Pagan

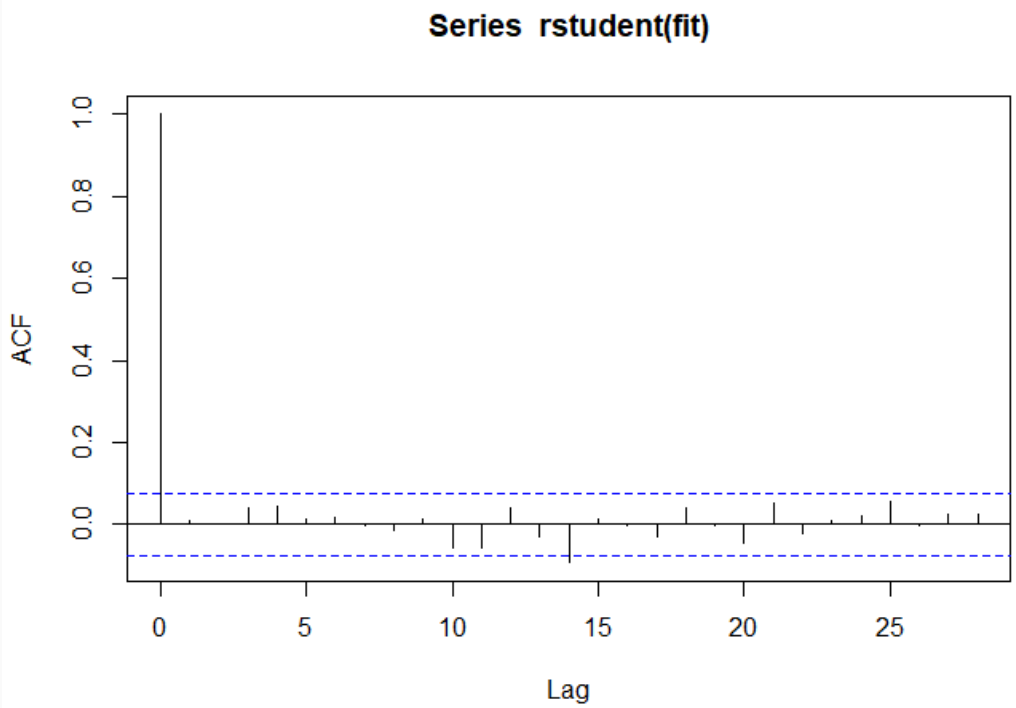
$$\begin{cases} H_0: \text{ Os erros são homoscedasticos} \\ H_1: \text{Os erros não são homoscedasticos.} \end{cases}$$

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 100.22, df = 18, p-value = 2.016e-13
```

Conforme o teste de Breusch-Pagan, rejeita-se H_0 devido p-valor = 2.0160966^{-13} < α = 0.01. Dessa forma, os erros são heteroscedasticos, não seguindo a suposição $[S2]$. O ideal para seria modelar a variância junto, já que existe muita diferença de manejo dos produtores entre os locais dos experimentos.

// Teste de Durbin-Watson

$$\begin{cases} H_0: \text{Não há autocorrelação} \\ H_1: \text{Há autocorrelação.} \end{cases}$$




```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.9708, p-value = 0.347
## alternative hypothesis: true autocorrelation is greater than 0
```

Conforme o teste de Durbin-Watson, não rejeita-se H_0 devido p-valor = 0.3469988 > α = 0.01. Ou seja, não existe multicolinealidade entre as variáveis explicativas

// Fatores de Inflação de Variância

$$\begin{cases} H_0 : \text{Não há multicolinearidade} \\ H_1 : \text{Há multicolinearidade.} \end{cases}$$

	GVIF	Df	GVIF^(1/(2*Df))
Terras	5.036604	1	2.244238
Ambiente	3.174420	1	1.781690
Cultura_Ant	54.890862	7	1.331220
P_base	2.649398	1	1.627697
N_base	3.077642	1	1.754321
GMR	1.113293	1	1.055127
Espacamento	2.360238	1	1.536307

	GVIF	Df	GVIF^(1/(2*Df))
Temperatura_Max	3.029015	1	1.740407
PH	2.649733	1	1.627800
`M.O.(%)`	2.279752	1	1.509885
mes	6.295974	2	1.584038

Interpretação:

vif maior que 10 indica multicolinearidade, vif próximo de 1 seria o ideal.

Dessa forma, todos os valores estão próximos de 1 indicando o indício de não multicolinearidade

// Teste Jarque-Bera

$$\begin{cases} H_0 : \text{Os erros possuem distribuição normal} \\ H_1 : \text{Os erros não possuem distribuição normal.} \end{cases}$$

```
##
##  Jarque Bera Test
##
## data:  resid(fit)
## X-squared = 7.3333, df = 2, p-value = 0.02556
```

Conforme o teste de Jarque-Bera, não rejeita-se H_0 devido ao p-valor = 0.0255623$ > = 0.01$. Ou seja, os erros possuem distribuição normal

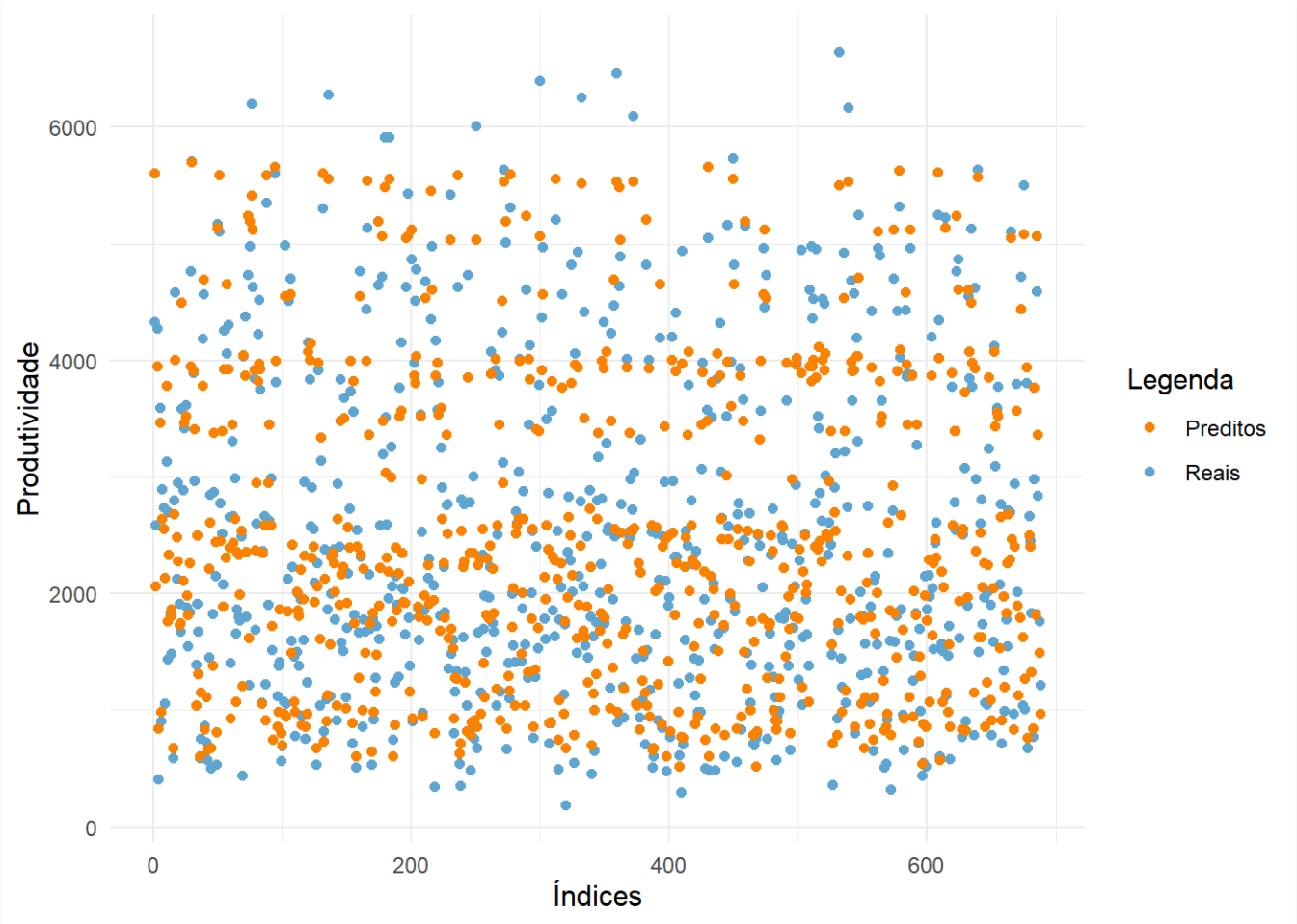
/ Predição

Dado que as suposições do modelo foram satisfeitas, podemos realizar inferências e predições sobre a variável resposta. Abaixo, é apresentado uma tabela com valores aleatórios,dentro do espaço amostral, para as variáveis significativas do modelo e assim, realizar a predição da produtividade média.

Novos Dados e Predição

Terras	Ambiente	Cultura_Ant	P_base	N_base	GMR	Espacamento	Temperatura_Max	PH	M.O. (%)	mes	Predicao
1	0	aveia	70	10	5.0	0.50	25	5	2	nov	387.4551
1	1	cevada	80	6	6.0	0.45	26	7	3	out	4494.2662
0	0	aveia e ervilhaca	75	7	5.5	0.45	26	8	4	dez	2012.5757

O gráfivo a seguir mostra a relação entre os valores reais de produtividade, com os valores preditos a partir do modelo de regressão.



Conclusão

De acordo com a análise de regressão, as principais variáveis que influenciam na produtividade da soja neste banco de dados são o tipo de terras e ambiente, o espaçamento entre linhas, a temperatura máxima durante o período de crescimento, o pH do solo e o teor de matéria orgânica. Além disso, práticas agrícolas como a escolha da cultura anterior e a quantidade de adubação com fósforo e nitrogênio também mostraram impacto significativo.

A análise diagnóstica do modelo identificou alguns pontos influentes e de alavancagem, mas estes não comprometeram significativamente a qualidade do ajuste. A verificação das suposições do modelo mostrou que a maioria foi atendida, exceto pela homoscedasticidade dos erros, indicando a presença de heteroscedasticidade. Isso sugere que este não é o melhor modelo, mas ainda oferece insights valiosos sobre os fatores que afetam a produtividade da soja.

Trabalhos futuros

- Modelar a variância
- Remover pelo menos uma safra, em locais que foi realizado os ensaios em mais de um ano, para assim, evitar possível dependência temporal.
- Usar todos os blocos invés da média deles
- Pensar em uma forma de definir todos os dias de plantio
- Usar as coordenadas geográficas dos local como variáveis