Modelagem via regressão linear do lucro de Startups

Caroline Cogo Carneosso

agosto 2022

Introdução

Startup é uma empresa inovadora com custos de manutenção muito baixos, mas que consegue crescer rapidamente e gerar lucros cada vez maiores. Além disso, possui um modelo de negócios repetível, escalável, em um cenário de incertezas e soluções a serem desenvolvidas.

Assim, a proposta do presente trabalho é propor um modelo de regressão linear que seja capaz de avaliar o efeito das covariáveis na média da variável y e por consequência ter uma boa predição dos valores de y. Para a validação deste modelo será utilizado critérios de seleção, gráficos para a análise de diagnóstico e influência, tais como: alavancagem, distância de Cook, envelope simulado, entre outras técnicas gráficas.

O banco de dados é referente a detalhes da receita de 50 Startups dos estados de New York, California e Florida, disponível na plataforma Kaggle.

Descrição da variáveis

Variável	Descrição
`$y$` profit	Lucro total da Startup
`$x_1$` adm	Gastos com Administração
`$x_2$` rdspend	Gastos com Pesquisa e Desenvolvimento
`$x_3$` mkt	Gastos com Marketing
`$x_4$` estado	Estado da Startup, New York, California ou Florida

Para fins de interpretação vamos considerar os valores em dólares. O banco de dados possui 50 observações e 5 variáveis. Além disso, foram criadas 2 variáveis dummies para a variável categórica estado, utilizando o estado da California como base, ou seja, foram acrescentadas as cováriaveis: estadoFlorida e estadoNew York.

Análise descritiva das variáveis.

	rdspend	adm	mkt	estado	profit
	Min. : 0	Min. : 51283	Min. : 0	California:17	Min. : 14681
	1st Qu.: 39936	1st Qu.:103731	1st Qu.:129300	Florida :16	1st Qu.: 90139
	Median : 73051	Median :122700	Median :212716	New York :17	Median :107978
	Mean : 73722	Mean :121345	Mean :211025	NA	Mean :112013
	3rd Qu.:101603	3rd Qu.:144842	3rd Qu.:299469	NA	3rd Qu.:139766
	Max. :165349	Max. :182646	Max. :471784	NA	Max. :192262

Correlação entre as variáveis.

	rdspend	adm	mkt	profit
rdspend	1.000	0.242	0.724	0.973
adm	0.242	1.000	-0.032	0.201
mkt	0.724	-0.032	1.000	0.748
profit	0.973	0.201	0.748	1.000

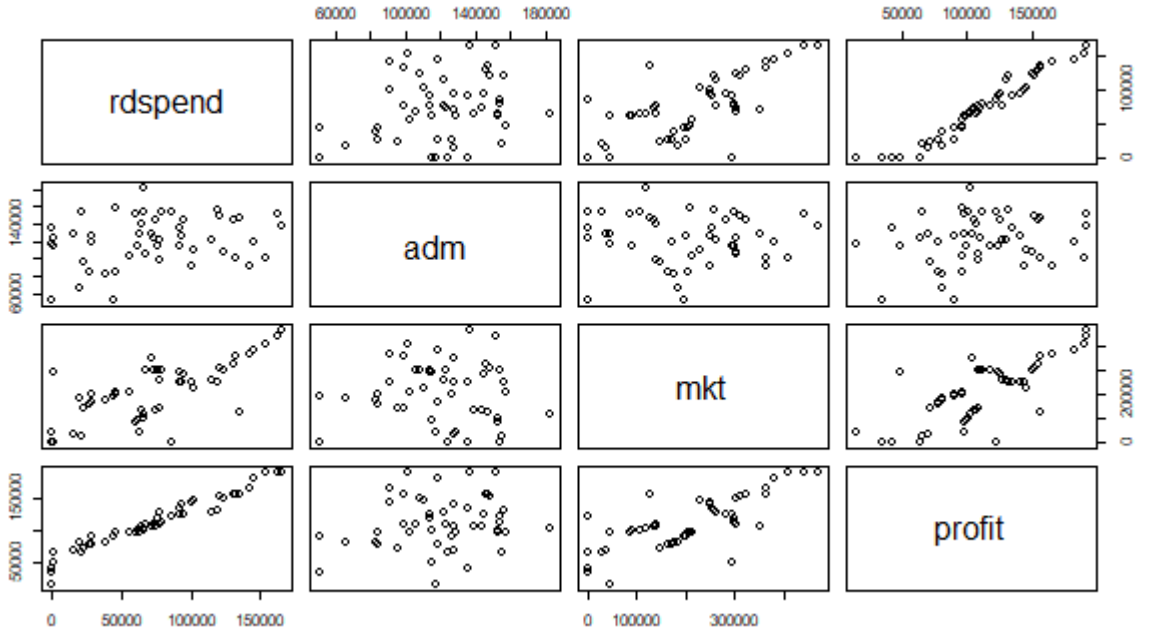


Gráfico de dispersões

Modelo inicial

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon.$$

No modelo inicial, apenas o intercepto e x_2 apresentaram significância. O modelo possui coeficiente de determinação (R2) de 0.951, e R2 ajustado $(\bar{R}2)$ de 0.945.

Coeficientes para o modelo inicial

	Estimate	Std. Error	t value	p.value
(Intercept)	50125.344	6884.820	7.281	0.001***
adm	-0.027	0.052	-0.517	0.608
rdspend	0.806	0.046	17.369	0.001***
mkt	0.027	0.017	1.574	0.123
estadoFlorida	198.789	3371.007	0.059	0.953
estadoNew York	-41.887	3256.039	-0.013	0.99

Note: Signif. codes 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Entretanto, ao aplicar a função $step$ que analisa através do critério de informação de Akaike (AIC) o melhor modelo a ser proposto, e retira possíveis covariáveis não explicativas, o método selecionou como significativa o intercepto e as covariáveis x_2 e x_3.

Modelo ajustado

Após sucessivas aplicações da função $step$, testando as combinações das covariáveis e diversas análises, de pontos influentes e gráficas, chegamos ao modelo ajustado apresentado.

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

O intercepto e todas as covariáveis foram significativas (x_2, x_3), para isso foi necessário retirar as observações **47 e 50** do banco de dados original.

Coeficientes para o modelo ajustado

	Estimate	Std. Error	t value	p.value
(Intercept)	50172.047	2333.087	21.50	0.001***
rdspend	0.751	0.039	19.43	0.001***
mkt	0.035	0.014	2.51	0.01*

Note: Signif. codes 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Análise de diagnóstico e influência

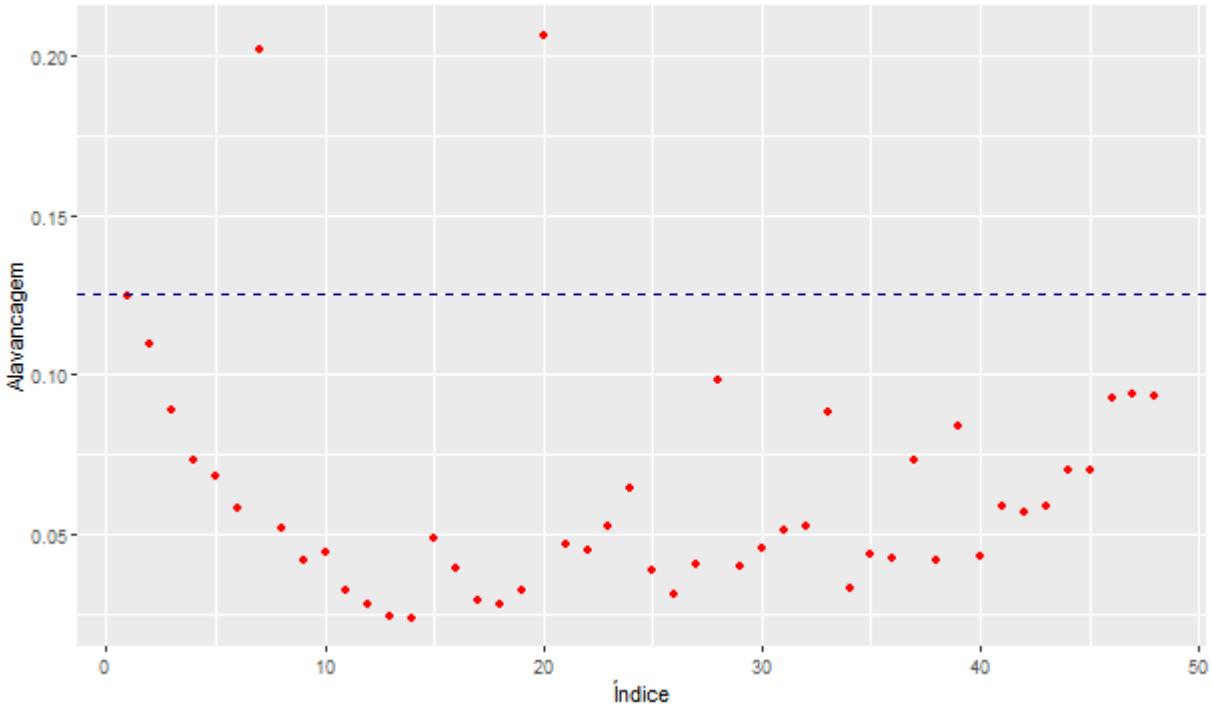


Gráfico para a alavancagem

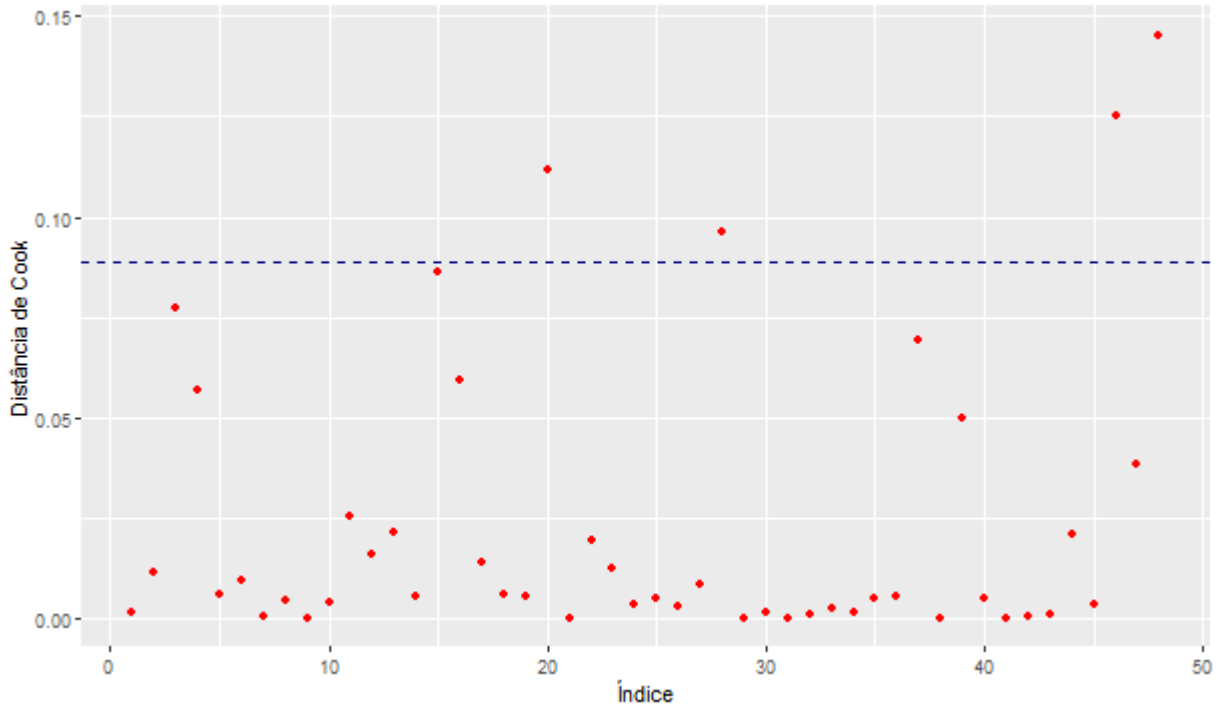
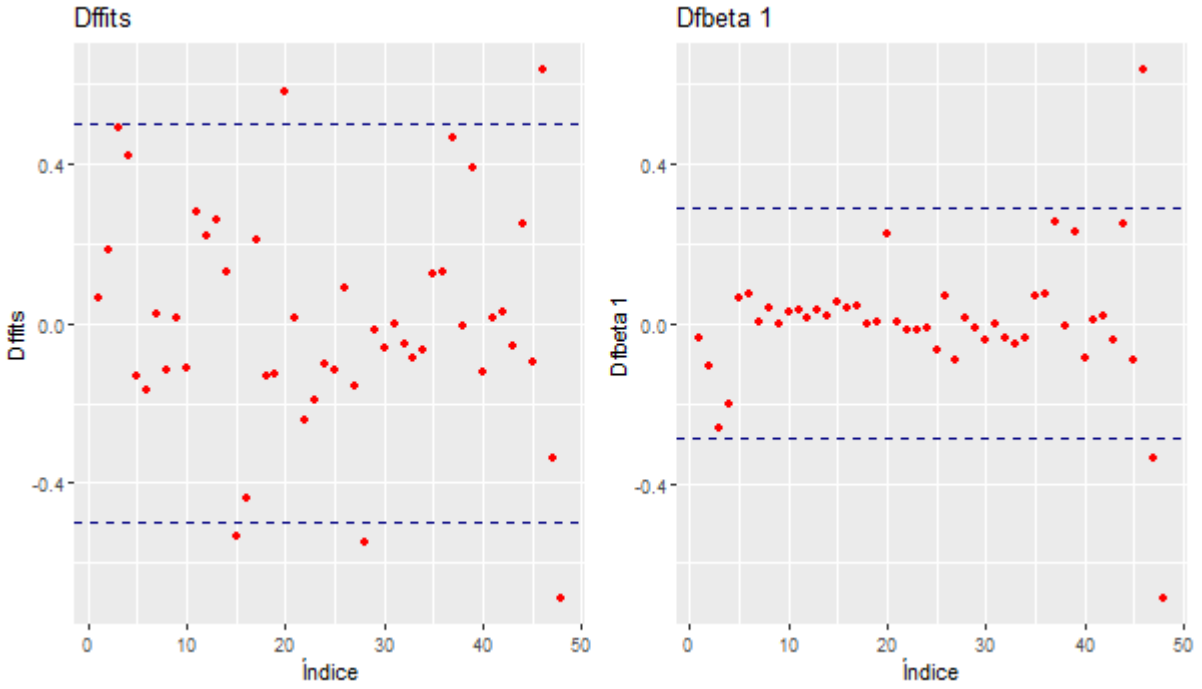
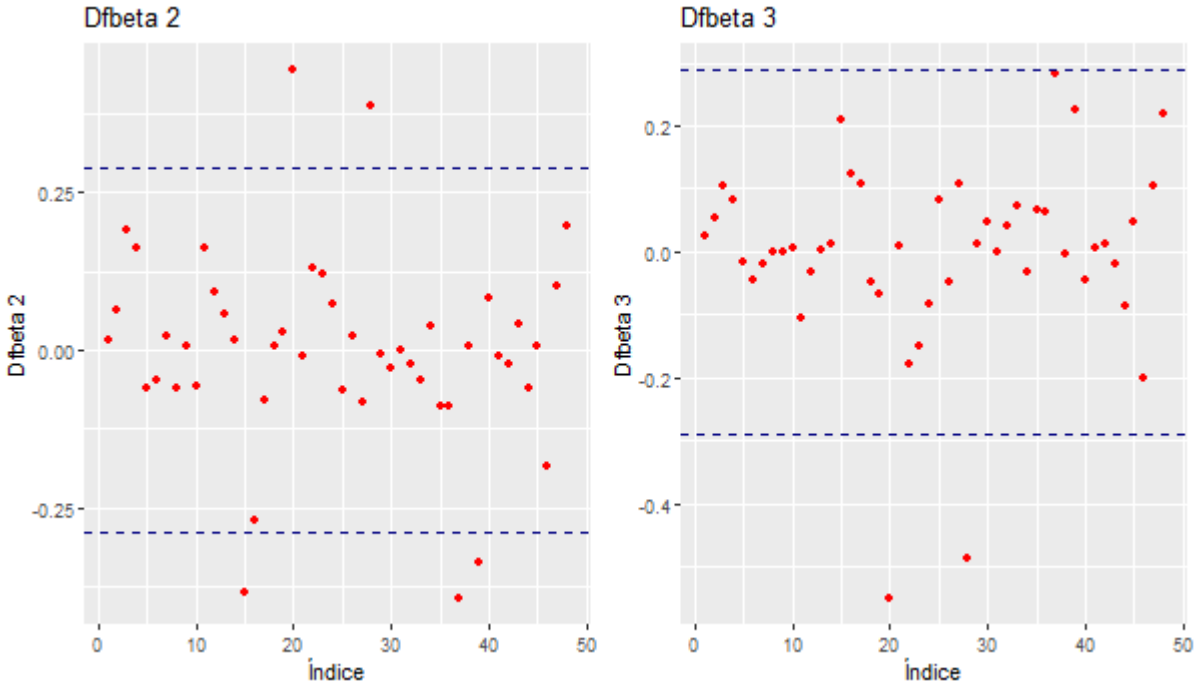


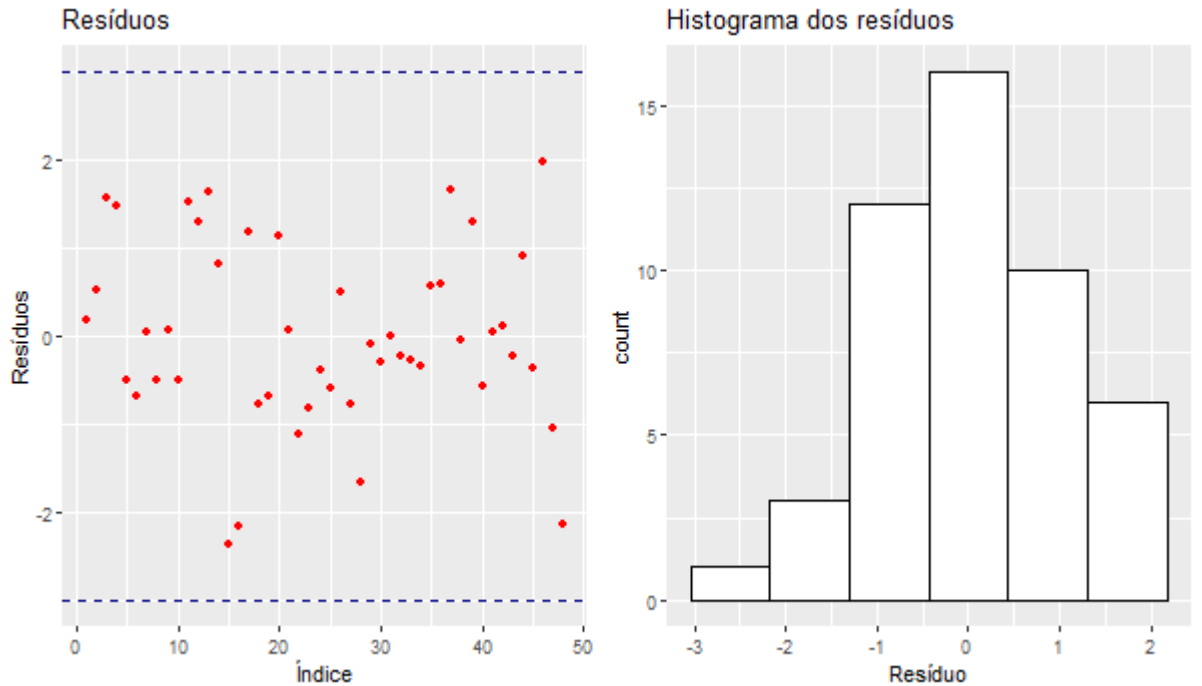
Gráfico para a Distância de Cook



Gráficos para o DFFIT e Dfbeta



Gráficos para os Dfbetas



Gráficos para o resíduo e histograma

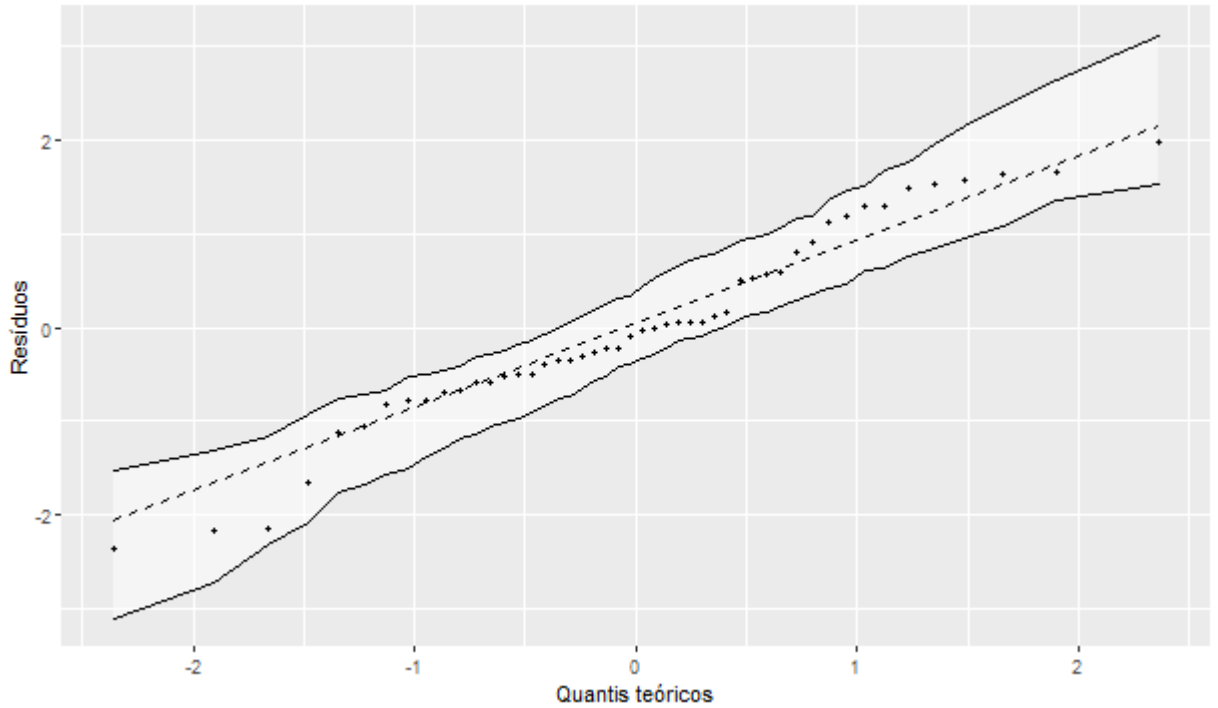


Gráfico de Envelope Simulado

Suposições do modelo

Para que o modelo seja validado é necessário confirmar as suposições abaixo através de testes.

- [S0] O modelo está corretamente especificado.
- [S1] A média dos erros é zero.
- [S2] Homoscedasticidade dos erros.
- [S3] Não há autocorrelação.
- [S4] Ausência de multicolinearidade.
- [S5] Normalidade dos erros.

Para testes de hipóteses, se $\alpha > p - valor$, então rejeita-se a hipótese nula **(H0)**.

Teste para a [S0]

Teste RESET de especificação sob **H0**: O modelo está corretamente especificado. Com p-valor igual a 0.24, ao nível de significância igual a $\alpha = 0.05$, não rejeitamos **H0**. Logo, não há evidências de incorreta especificação do modelo.

Teste para a [S1]

Teste t para a média dos errros sob **H0**: média dos erros é igual a zero. Com p-valor igual a 0.996, ao nível de significância igual a $\alpha = 0.05$, conclui-se que não rejeitamos **H0**. Logo, a média dos erros é igual a zero.

Teste para a [S2]

Teste de Bressch-Pagan (Koenker) de Heteroscedasticidade sob H0: erros são homoscedásticos. Com p-valor igual a 0.999, ao nível de significância igual a $\alpha = 0.05$, conclui-se que não rejeitamos **H0**. Logo, os erros são homoscedásticos.

Teste para a [S3]

Teste de Durbin-Watson de autocorrelação sob **H0**: não há autocorrelação. Com p-valor igual a 0.047, ao nível de significância igual a $\alpha = 0.05$, conclui-se que rejeitamos **H0**.

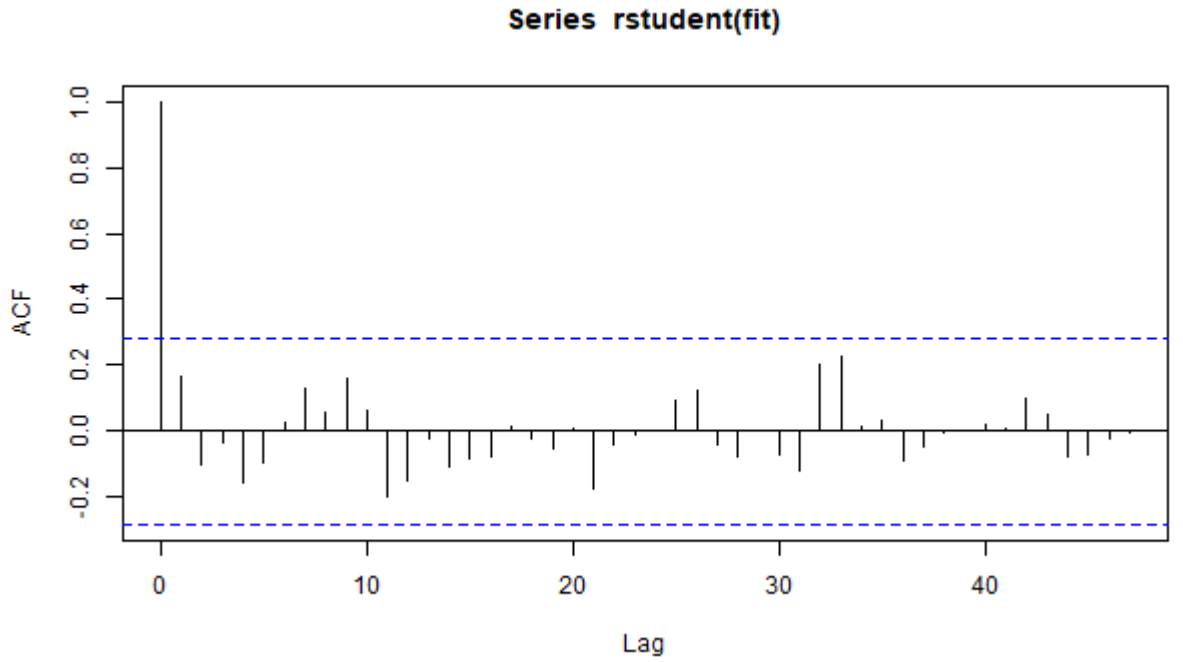


Gráfico da função de autocorrelação

Teste para a [S4]

Usa-se Fatores de Inflação de Variância (VIF) para detectar multicolinearidade. Em que, VIF > 10 indica multicolinearidade e VIF=1 seria o ideal.

Fatores de Inflação de Variância para as variáveis do modelo ajustado.

	x
rdspend	2.38
mkt	2.38

Percebe-se pela Tabela, que o valor está próximo a 1, para x_2 e x_3. Logo, não há multicolinearidade.

Teste para a [S5]

Teste Jarque-Bera de Normalidade, **H0**: Os erros possuem distribuição normal. Com p-valor igual a 0.882, ao nível de significância igual a $\alpha = 0.05$, conclui-se que não rejeitamos **H0**. Logo, não existem indícios de não normalidade dos erros.

Modelo final

O modelo apresentou $R2 = 0.96$, ou seja, 96% da variação da média do lucro das Startups é explicada por x_2 e x_3. Além disso, o critério de seleção do modelo é de $(\bar{R}2)$ ajustado igual a 0.959.

$$y = 50172.0465 + 0.7512x_2 + 0.0353x_3.$$

Nota-se que as covariáveis influenciam positivamente na média de y, e o intercepto também. Para a covariável x_2, a cada 1 dólar gastos com pesquisa e desenvolvimento, adiciona-se 0.75 dólares no lucro da Startup, para a covariável x_3, a cada 1 dólar gastos com marketing, adiciona-se 0.03 doláres no lucro da Startup.

Obrigada!