Modelagem via regressão linear do lucro de Startups

Caroline Cogo Carneosso*

agosto 2022

Sumário

*carolcogo808@gmail.com

1 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 definir um modelo de regressão linear que seja capaz de predizer a variável y, e quanto as covariáveis influenciam na média 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 e pode ser acessado clicando aqui.

Tabela 1: Descrição da váriaveis

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, que estão descritas na Tabela 1. Além disso, foram criadas 2 variáveis dummies para a variável categórica estado, utilizando o estado da California como base, ou seja, foram acrescentadas as covariáveis: estadoFlorida e estadoNew York.

2 Análise descritiva

Podemos avaliar pela Tabela 2, um resumo das variáveis, com as medidas descritivas, contendo o valor mínimo, 1º quantil, mediana, valor médio, 3º quantil e valor máximo. É importante ressaltar, que a variável de desfecho lucro, para as 50 startups do banco de dados, apresentou menor lucro de 14 681 dólares e lucro máximo de 192 262 dólares.

Tabela 2: 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

É importante examinar a correlação entre as covariáveis, pois devemos ter uma correlação "aceitável" entre a variável resposta e as covariáveis, o que entretanto, não pode acontecer entre as covariáveis, pois afeta negativamente o método de mínimos quadrados ordinários.

Tabela 3: 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

Notamos pela Figura 1 e Tabela 3 que Gastos com Marketing (mkt) e Gastos com Pesquisa e Desenvolvimento (rdspend) possuem uma correlação alta o que pode afetar a suposição [S4]. Além disso, existe uma correlação alta e positiva de 0.97 entre o lucro (profit) e o Gastos com Pesquisa e Desenvolvimento (rdspend), ou seja, quanto maior o gasto em pesquisa e desenvolvimento, maior será o lucro.

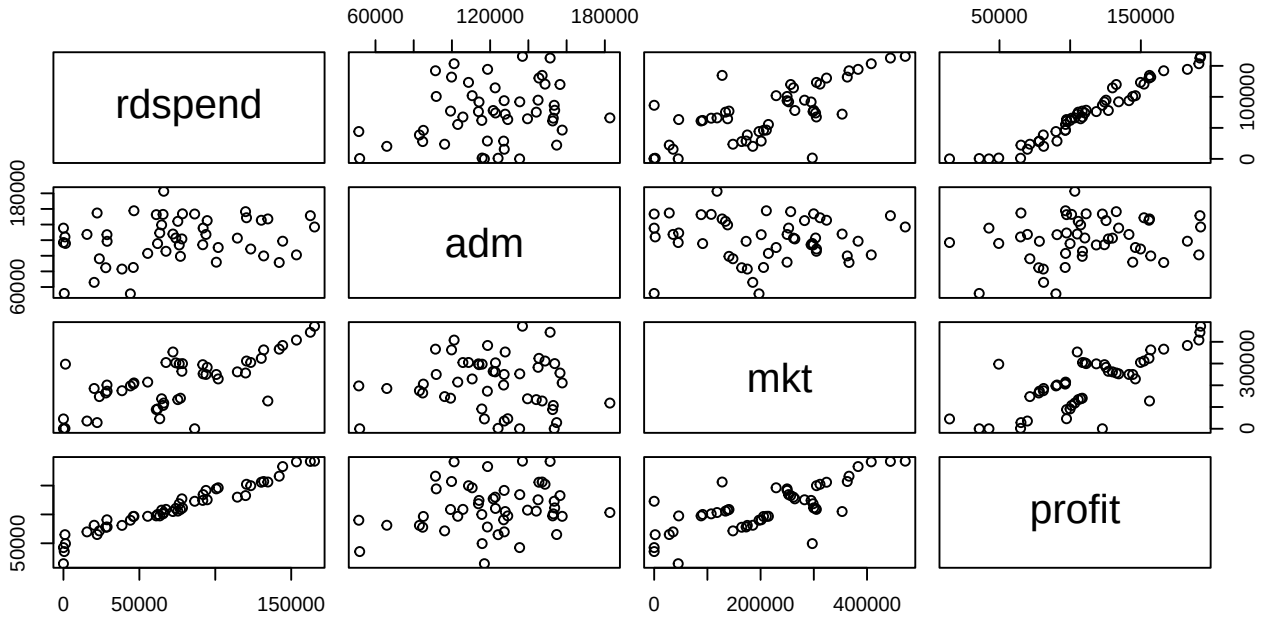

Figura 1: Gráfico de dispersões

3 Modelo inicial

Agora que já fizemos uma análise inicial das variáveis do estudo, apresenta-se o modelo inicial abaixo, contendo todas as variáveis.

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

Em que β_0 é o intercepto do modelo, y é a variável resposta, lucro da Startup, e o vetor de covariáveis $(x_1, x_2, x_3, x_4)^T$, foram descritas na Tabela 1.

No modelo inicial, através do teste t, apenas o intercepto e x_2 possui significância a um nível de 5%, como pode ser visto na Tabela 4. O modelo apresenta coeficiente de determinação (R2) de 0.951, e R2 ajustado ($\bar{R}2$) de 0.945.

Tabela 4: 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 o método de Stepwise, através da função *step* do software *R* 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.

4 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 abaixo.

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

Na Tabela 5, 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.

Tabela 5: 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

5 Análise de diagnóstico e influência

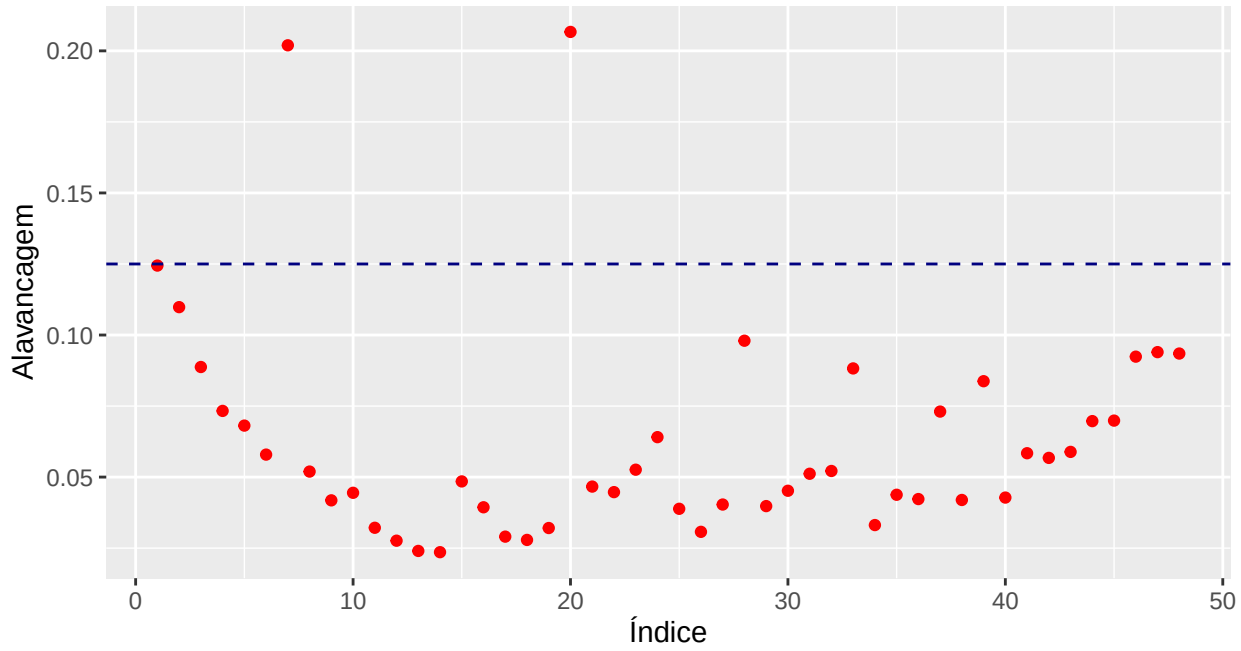

Figura 2: Gráfico para a alavancagem

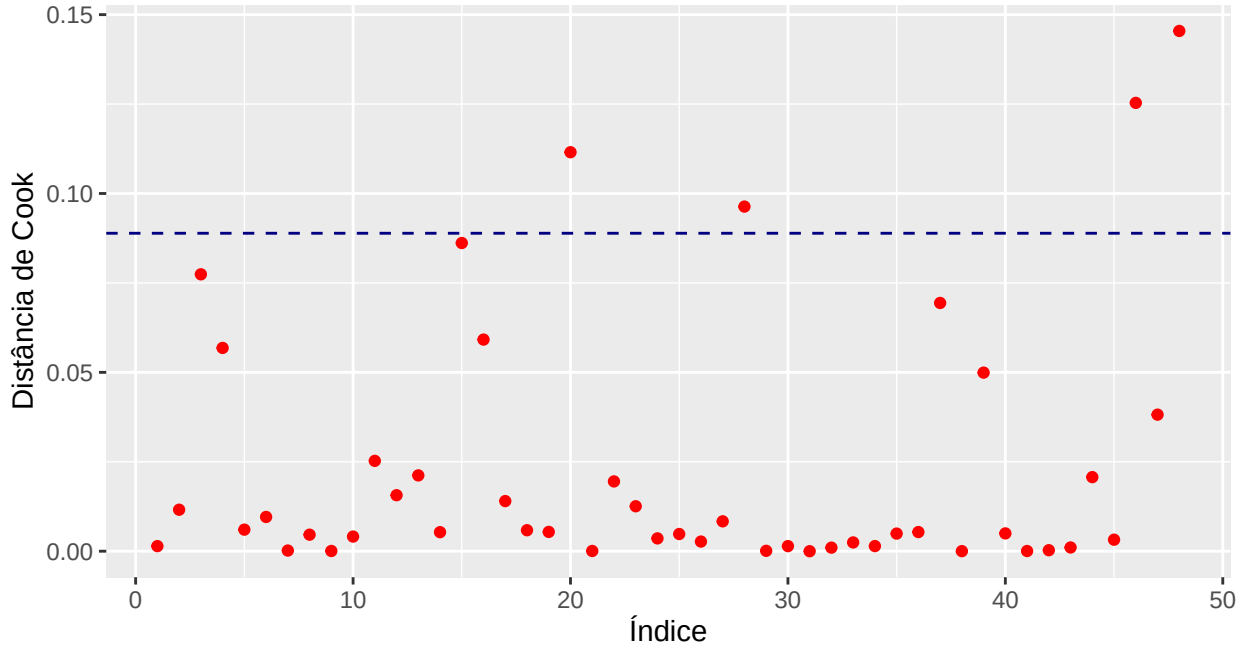

Figura 3: Gráfico para a Distância de Cook

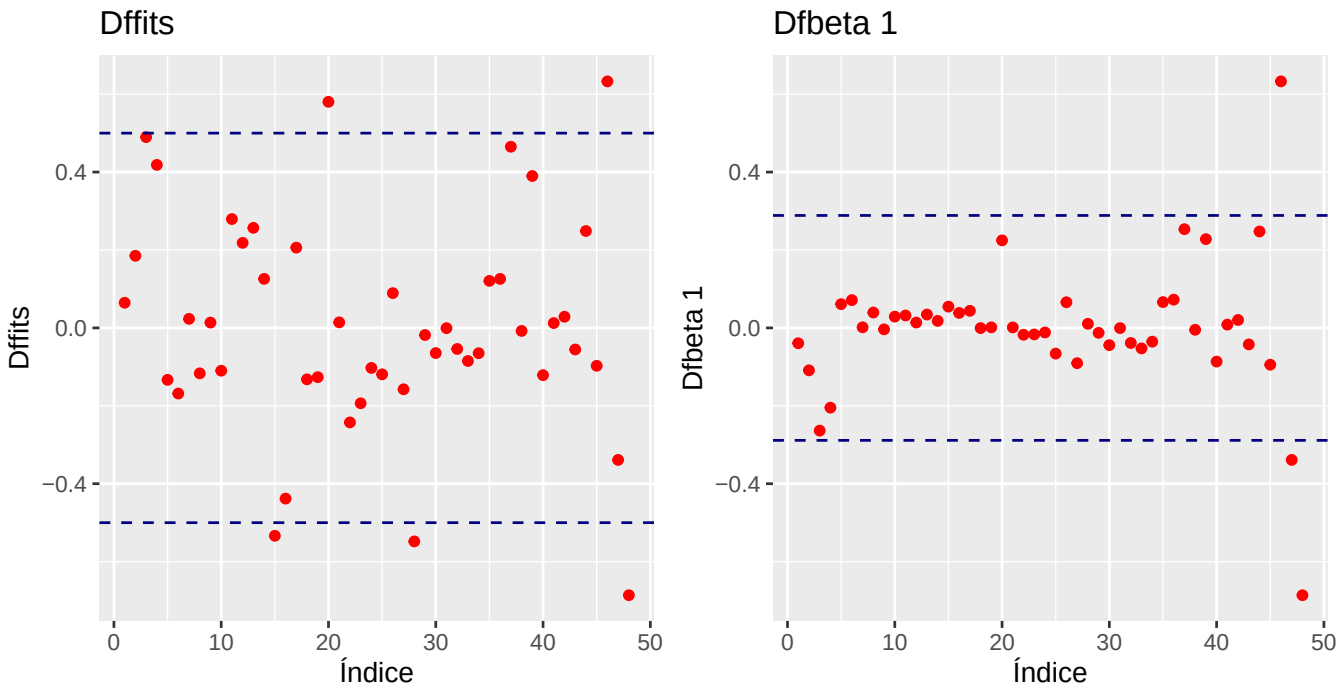

Figura 4: Gráficos para o DFFIT e Dfbeta

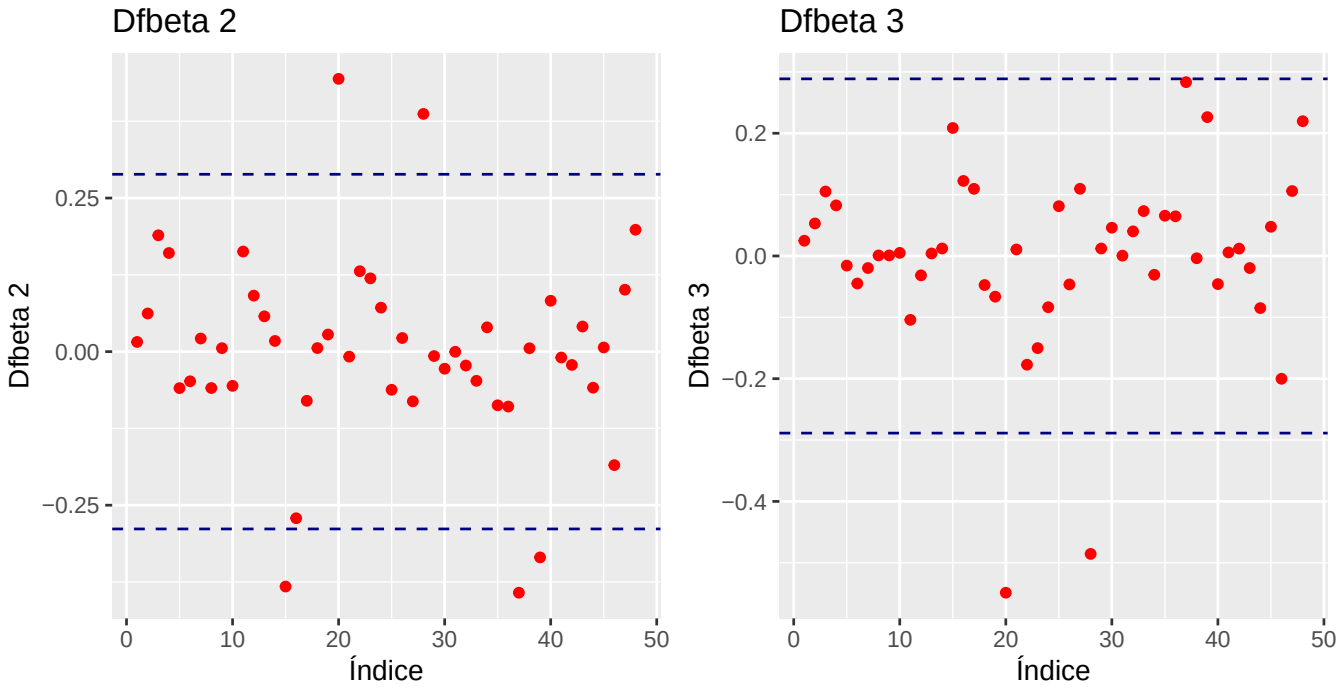

Figura 5: Gráficos para os Dfbetas

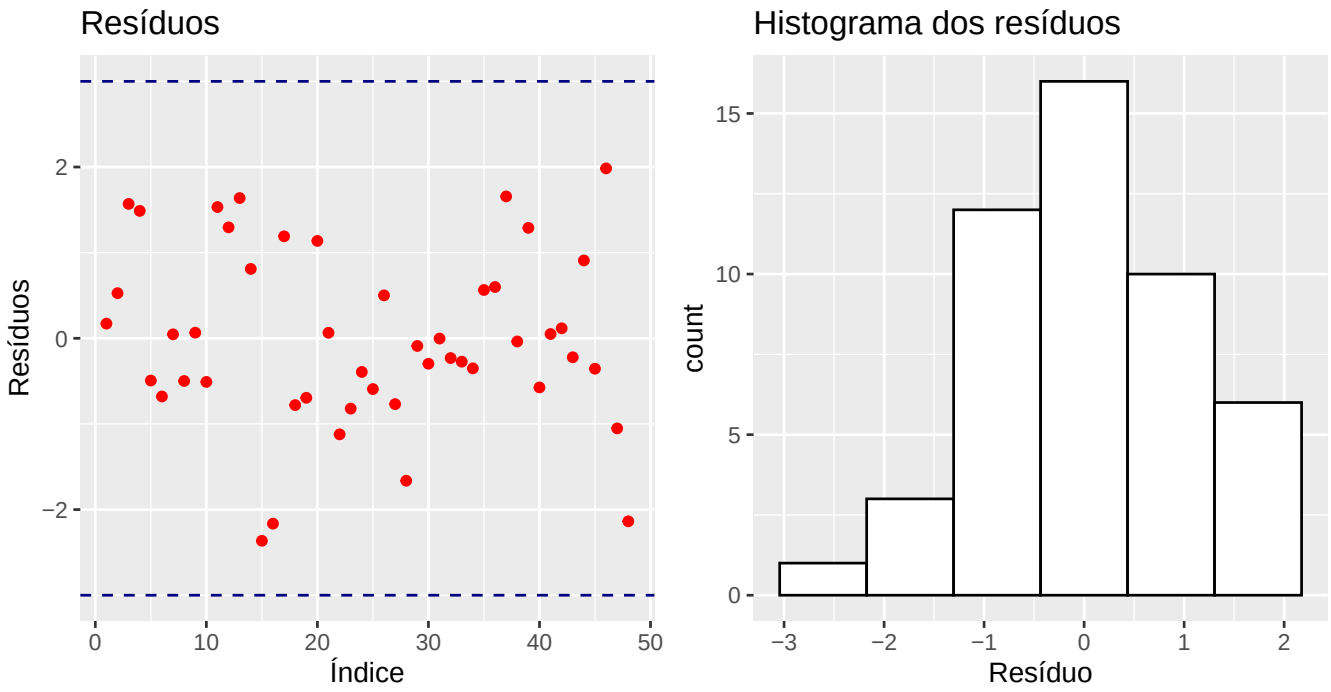

Figura 6: Gráficos para o resíduo e histograma

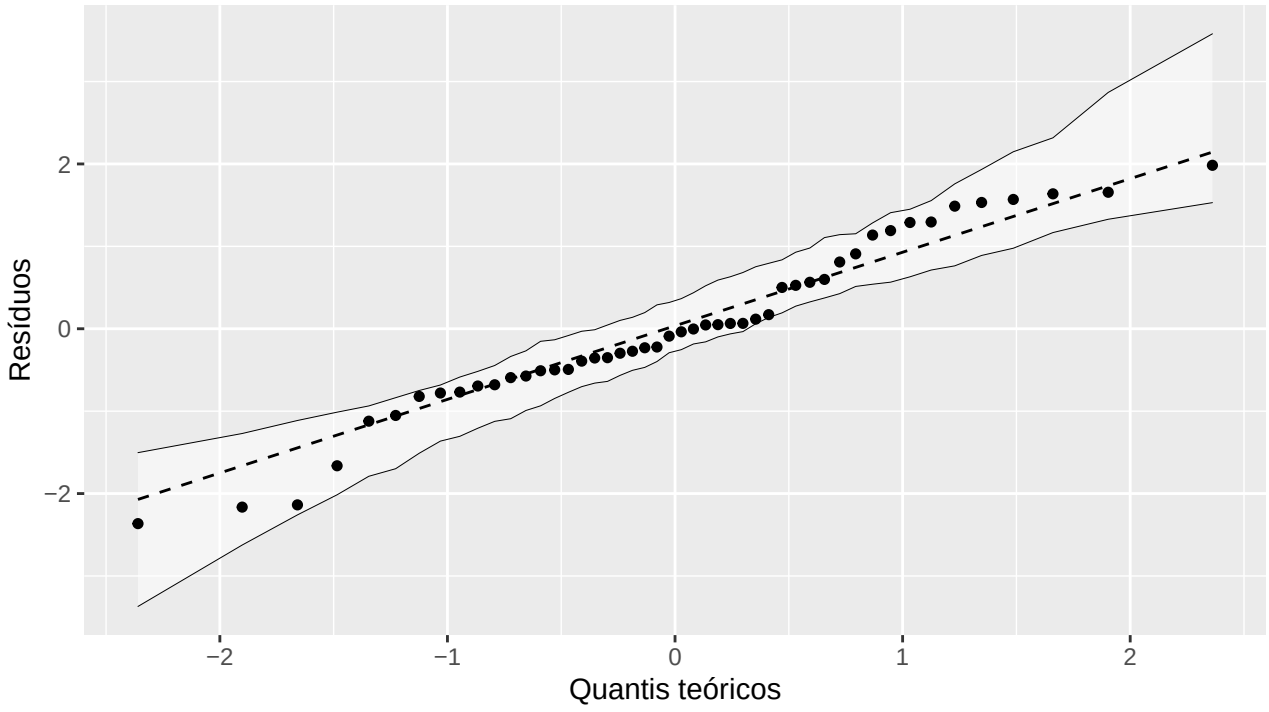

Figura 7: Gráfico de Envelope Simulado

Previamente, foi realizada a análise de influência, foram retiradas do banco de dados as observações 47 e 50, que se mostraram discreprantes das demais, e estavam interferindo na modelagem.

Na Figura 2, temos o gráfico das medidas de alavancagem, que informam se uma observação é discrepante em termos de covariável, nota-se que apenas duas observações estão um pouco fora dos limites pré-estabelecidos.

Através da Figura 3, observa-se o gráfico com a distância de Cook, que fornece a influência de cada observação i sobre todos os n valores ajustados, há alguns pontos fora do limite, porém não são discrepantes dos demais.

Por meio da Figura 4, temos o gráfico dos Dffits que considera o grau de influência que a observação i tem sobre o valor seu próprio valor ajustado $\hat{y}_i$. Na Figura 4 e 5, visualizamos os gráficos para os Dfbetas, que medem a influência da observação i sob as estimativas de cada β. Em ambas as figuras, não percebemos nenhum ponto claramente influente.

Ao visualizar, a Figura 6, temos o gráfico dos resíduos, onde percebe-se que todos as observações estão dentro do limite de 3 desvios padrões, e também o histograma, onde nota-se que os resíduos se assemelham a uma distribuição normal.

Na Figura 7, temos o envelope simulado baseado nos resíduos studentizados, com todas as observações dentro das bandas de confiança, o que sinaliza que a distribuição normal é adequada para o modelo.

6 Suposições do modelo

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

- [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**).

6.1 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.

6.2 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.

6.3 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.

6.4 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**. Logo, através do teste, concluimos que há autocorrelação. Entretanto, pode se ver através do gráfico da Figura 8, que não existe autocorrelação já que todos os lags são não significativos, ou seja, dentro das bandas de confiança. Assim sendo, vamos considerar a não existência de correlação e a verificação da S3.

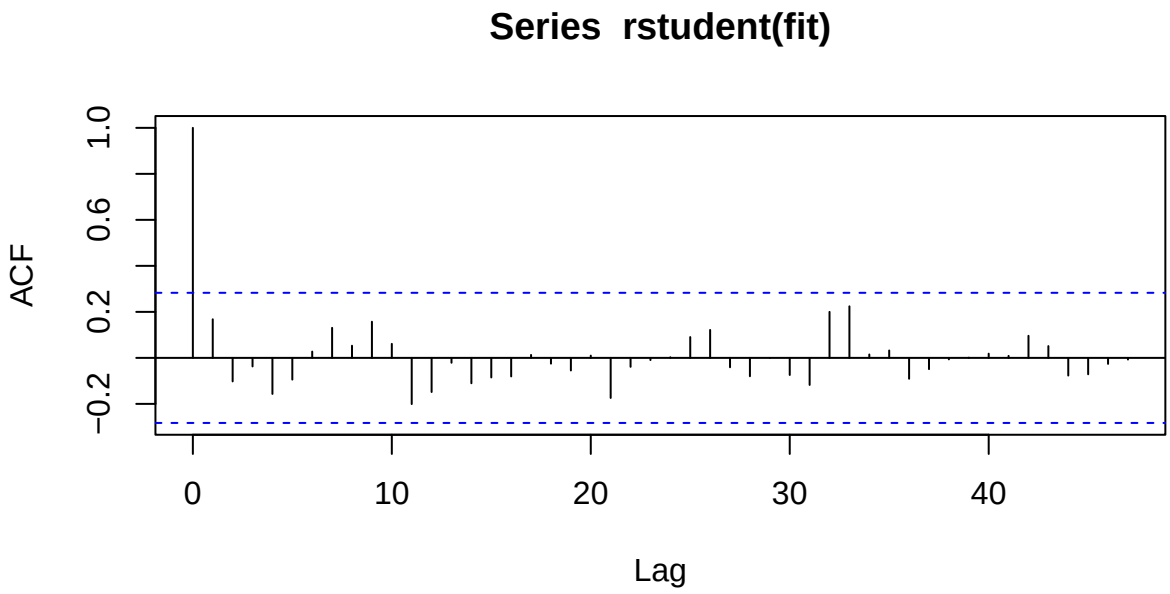

Figura 8: Gráfico da função de autocorrelação

6.5 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.

Tabela 6: Fatores de Inflação de Variância para as variáveis do modelo ajustado.

	x
rdspend	2.38
mkt	2.38

Percebe-se pela Tabela 6, que o valor está próximo a 1, para x_2 e x_3. Logo, não há multicolinearidade.

6.6 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.

7 Modelo final

Agora com o modelo checado, com boas evidências de que as suposições estão satisfeitas, é possivel fazer interpretações. O modelo apresentou $R2 = 0,96$, cerca de 96% da variação de y, é explicado pelas covariáveis, ou seja, 96% da variação da média do lucro das Startups é explicada por x_2 e x_3, gastos com pesquisa e desenvolvimento e gastos em marketing, respectivamente. 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.

8 Conclusão

Portanto, O estudo propôs um modelo de regressãao linear para o banco de dados de 50 Startups, contendo como variável resposta o Lucro e como covariáveis Gastos com Administração, Gastos com Pesquisa e Desenvolvimento, Gastos com Marketing e Estados. Do modelo inicial foram retiradas as covariáveis Gastos com Administração e as dummies relacionadas aos estados, assim como as observações 47 e 50, que apresentaram forte influência durante as análises. Aproximadamente 96% da da variação de y, é explicado pelas covariáveis, o que indica um bom ajuste do modelo. Também pode-se concluir que as covariáveis influenciam positivamente na média de y, assim como o intercepto.

Bibliografia

[1] R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/.

[2] Kaggle. Startup - multiple linear regression [Internet]. 2022. Available from: https://www.kaggle.com/datasets/karthickveerakumar/startup-logistic-regression.