

Análise de Resíduos em Modelos de Regressão Linear

Testando Soluções com R

Francisco Piccolo

2019-09-05

O modelo de regressão linear é bastante usado na predição de variáveis contínuas, onde há uma ou mais variáveis independentes buscando mapear o comportamento de uma variável dependente. O modelo é bastante simples e lembro ser um dos primeiros a ser ensinado nas aulas de econometria. Porém, apesar de sua simplicidade, é preciso se atentar a alguns detalhes sobre suas premissas para que os resultados deste modelo possam ser usados para a tomada de decisão.

Abaixo vou listar algumas das premissas da regressão linear:

- i) Ausência de multicolinearidade entre as variáveis independentes.
- ii) Ausência de autocorrelação na variável dependente.
- iii) Absence of pattern on the behavior of the model residuals, in other words, absence of heteroscedasticity.
- iv) Ausência de padrão no comportamento dos resíduos do modelo, ou seja, ausência de heterocedasticidade.
- v) Resíduos se distribuem de acordo com uma distribuição normal.

Tendo estas premissas atendidas, o modelo pode gerar conclusões confiáveis. Neste post eu vou desenvolver um modelo de regressão linear para testar as premissas **iii** e **iv** que tratam dos resíduos, para ver alguns casos práticos. O código usado neste post está nesta [pasta](#) do meu Github.

Exemplo prático: Elasticidade preço x oferta na produção de cana de açúcar

Este exemplo foi passado na minha aula de econometria em 2017. Na época o exercício foi realizado com o software **EViews**, por sorte eu guardei os dados e agora posso refazer o problema com mais facilidade com o uso do R. O dataset pode ser visualizado neste [link](#).

No exercício, temos a variável independente (X) sendo o preço da cana de açúcar e a variável dependente (Y) sendo a área plantada de cana de açúcar (representando uma proxy para a oferta do produto). O objetivo deste modelo é tentar quantificar a elasticidade da oferta em função do preço, ou seja, quantificar quão sensível é a oferta da cana de açúcar quando ocorrem variações em seu preço.

Abaixo há uma amostragem do dataset.

Table 1: Amostra do Dataset

	period	area	price
	15	15	125
	3	3	42
	27	27	97
	2	2	71
	11	11	88

O modelo de regressão linear para este cenário será desenvolvido de acordo com a fórmula abaixo:

$$\ln Y_t = \beta_0 + \beta_1(\ln X_t) + \mu_t$$

Onde:

Y_t = Área plantada após a transformação com log natural (e)

X_t = Preço da cana de açúcar também após a transformação com log natural

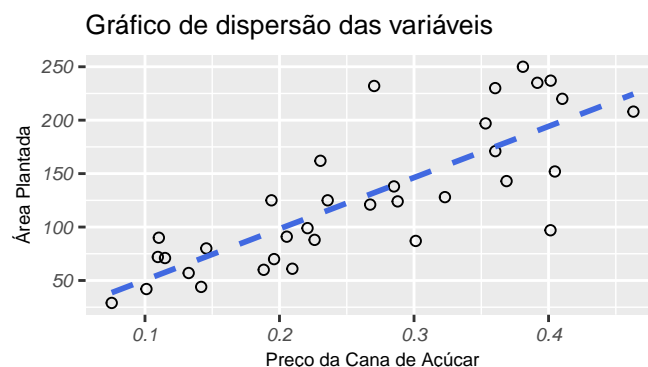
β_0 = Intercepto

β_1 = Inclinação

μ_t = Resíduos

Os dados precisam ter a aplicação do log natural, pois esta transformação faz com que as variações entre os períodos possam ser interpretadas como variações percentuais, e isso é necessário por conta de que a elasticidade é quantificada em termos percentuais. Esta característica ocorre apenas na transformação com logaritmo natural, se a transformação fosse feita com outros logs, a interpretação (de variações percentuais) não seria válida.

O gráfico abaixo irá mostrar o comportamento das variáveis do dataset, bem como a curva de regressão linear, antes de aplicar a transformação log natural.



Podemos ver que há uma relação entre o preço do produto e sua oferta (área plantada). Agora vamos aplicar o log natural no modelo de regressão. Para isso, o R nos fornece duas opções:

- Ajustar as variáveis no dataset e construir o modelo usando as variáveis ajustadas.
- Construir o modelo e indicar “dentro dele” que é necessário fazer a transformação antes de computar os resultados.

Vamos ver na prática como cada opção pode ser usada. O resultado final será idêntico.

Primeiro vou criar duas variáveis com os resultados dos dois métodos:

Com as duas variáveis criadas, vamos criar uma tabela comparando os principais resultados dos modelos:

Table 2: Comparativo dos Resultados

	X1º.Método	X2º.Método
(Intercept)	6.1113284	6.1113284
price_log	0.9705823	0.9705823

Conforme indicado, ambos os métodos geram o mesmo valor. Eu prefiro o segundo, que exige menos linhas de código. Com base nos coeficientes estimados, temos a seguinte equação:

$$\hat{Y} = 1.6416 + 0.9706X_1 + \mu$$

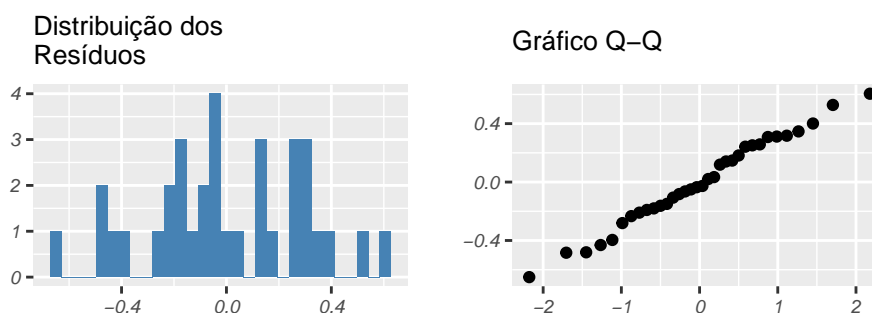
O resultado é estatisticamente significativo, visto que tanto o intercepto quanto a inclinação apresentam um valor-p baixo. Veja abaixo estes valores bem como o R^2 .

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.111	0.1686	36.25	1.468e-27
log(price)	0.9706	0.1106	8.773	5.031e-10

Table 4: Fitting linear model: log(area) ~ log(price)

Observations	Residual Std. Error	R^2	Adjusted R^2
34	0.3088	0.7063	0.6972

Embora o modelo consiga explicar ~70% da variação na variável dependente, é preciso analisar os resíduos gerados para poder ter confiança no resultado e fazer projeções (objetivo principal). No gráfico abaixo, vamos ver como se distribui os resíduos do modelo em um Histograma e Gráfico QQ (quantile-quantile)



Ambos os gráficos indicam distribuição normal dos resíduos. Apesar desta forma visual ser recomendada, as vezes ela não é suficiente, casos em que o pesquisador precisará usar métodos mais formais para gerar uma conclusão. Para validação da independência no comportamento dos resíduos pode-se usar o teste **Durbin Watson**. O código abaixo realiza este teste.

Table 5: Durbin-Watson test: `df %>% lm(formula = log(area) ~ log(price))`

Test statistic	P value	Alternative hypothesis
1.291	0.009801 * *	true autocorrelation is greater than 0

O resultado do teste foi de 1.2912, mas apenas com este valor não é possível fazer uma conclusão. Em conjunto com este valor é preciso saber os valores limiares **DL** e **DU**, que podem ser encontrados nesta [tabela](#). Para encontrar os valores com nesta tabela, basta saber o número de observações no dataset (i.e. 33), o nível de significância do teste (i.e. 0.05) e os graus de liberdade (i.e. 1). Com isso, tem-se:

DL = 1.35

DU = 1.49

Tendo DW igual a 1.2912, acima de 0 e abaixo de DL, pode-se concluir que os resíduos são independentes.

Com isso, podemos concluir que de fato o modelo é confiável para realizar projeções, pois tanto os gráficos quanto o teste formal indicam que as premissas (iii) e (iv) estão sendo atendidas. Desta forma, podemos concluir que há elasticidade na oferta de cana de açúcar com relação ao seu preço.