ANÁLISE DOS RESÍDUOS

Na análise de regressão linear, assumimos que os erros $E_1, E_2, \ldots, E_n$ satisfazem os seguintes pressupostos:

- seguem uma distribuição normal;
- têm média zero;
- têm variância σ^2 constante (homocedasticidade);
- são independentes.

A verificação das hipóteses é fundamental, visto que toda a inferência estatística no modelo de regressão linear (testes de hipóteses) se baseia nesses pressupostos. Nesse sentido, se houver violação dos mesmos, a utilização do modelo deve ser posta em causa.

A análise dos resíduos é uma ferramenta popular para detectar violações de tais pressupostos.

Recorda-se que o i-ésimo **resíduo d_i** é a diferença entre o valor observado y_i e o valor estimado $\hat{y}_i = \hat{\mu}_{Y/x_i}$ dado pela equação de regressão linear estimada.

NORMALIDADE DOS E_i's

O pressuposto de normalidade pode ser testado recorrendo a testes de ajustamento tais como o Teste Kolmogorov-Smirnov ou o Teste da Normalidade de Lilliefors, que serão abordados posteriormente no capítulo IV.

Essa condição também pode ser verificada usando um **gráfico de probabilidade normal** (Normal Probability Plot).

Existem dois tipos de gráficos de probabilidade normal:

- **1º tipo**: representa a probabilidade acumulada que seria de esperar se a distribuição fosse normal, em função da probabilidade observada acumulada dos erros (**Normal P-P Plot**);

- **2º tipo**: representa o quantil de probabilidade esperado se a distribuição fosse normal em função dos resíduos (**Normal Q-Q Plot**).

Para produzir estes gráficos, começa-se por **estandardizar os resíduos** de forma a terem um desvio padrão unitário:

$$d_i' = \frac{d_i - 0}{S} \qquad \text{onde} \qquad S^2 = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n} d_i^{\,2}}{n-k-1},$$

e **ordenam-se** por ordem crescente.

Em função do tipo de gráfico, calcula-se:

☆ **Normal P-P Plot** → o valor da função de distribuição para cada resíduo estandardizado, assumindo que têm uma distribuição normal; estes valores são representados no eixo das ordenadas

→ a probabilidade observada acumulada usando a fórmula

$$\frac{i-0.5}{n};$$

estes valores representam-se no eixo das abcissas.

☆ **Normal Q-Q Plot** → os quantis de probabilidade esperados, ou seja, os z_i tais que $P(Z < z_i) = \frac{i - 0.5}{n}$; estes valores são representados no eixo das ordenadas

→ no eixo das abcissas representam-se os resíduos estandardizados.

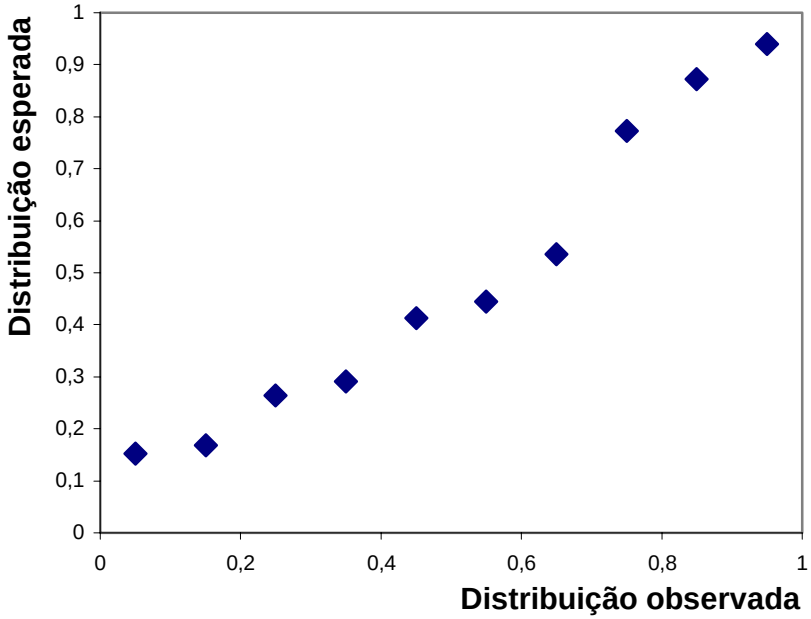
Se os erros possuírem distribuição Normal, todos os pontos dos gráficos devem posicionarem-se mais ou menos sobre uma recta.

Exemplo:

Para o exemplo que temos vindo a estudar, apresentam-se os resíduos estandardizados (ordenados) e os cálculos necessários para construir os gráficos de probabilidade normal.

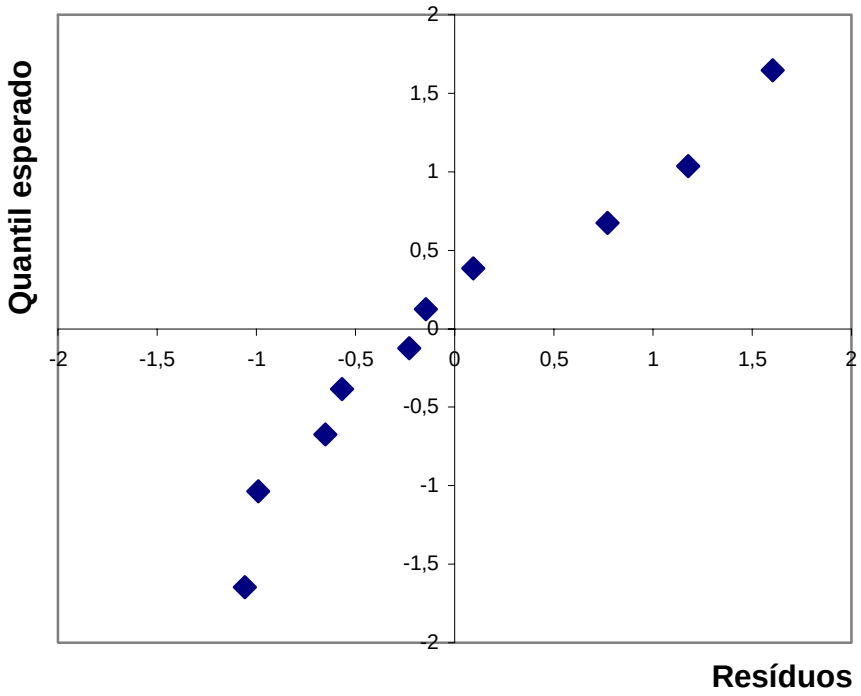
i	Resíduos d_i	Resíduos estandardizados $d_i' = d_i / S$ (abcissa Q-Q Plot)	Distribuição observada $\frac{i-0.5}{10}$ (abcissas P-P Plot)	Quantil esperado z_i: $P(Z < z_i) = \frac{i-0.5}{10}$ (ordenadas Q-Q Plot)	Distribuição esperada (ordenadas P-P Plot)
1	-1,05932	-1,02456	0,05	-1,644853627	0,152786
2	-0,99153	-0,95899	0,15	-1,036433389	0,168781
3	-0,65254	-0,63113	0,25	-0,67448975	0,263979
4	-0,5678	-0,54917	0,35	-0,385320466	0,291445
5	-0,22881	-0,2213	0,45	-0,125661347	0,412429
6	-0,14407	-0,13934	0,55	0,125661347	0,44459
7	0,09322	0,090161	0,65	0,385320466	0,53592
8	0,77119	0,745883	0,75	0,67448975	0,772131
9	1,17797	1,139315	0,85	1,036433389	0,872714
10	1,60169	1,549131	0,95	1,644853627	0,939325

Normal P-P Plot



Os pontos do gráfico tendem a concentrar-se em torno da recta de declive 1 que passa na origem, o que dá evidência de que a distribuição dos erros é normal.

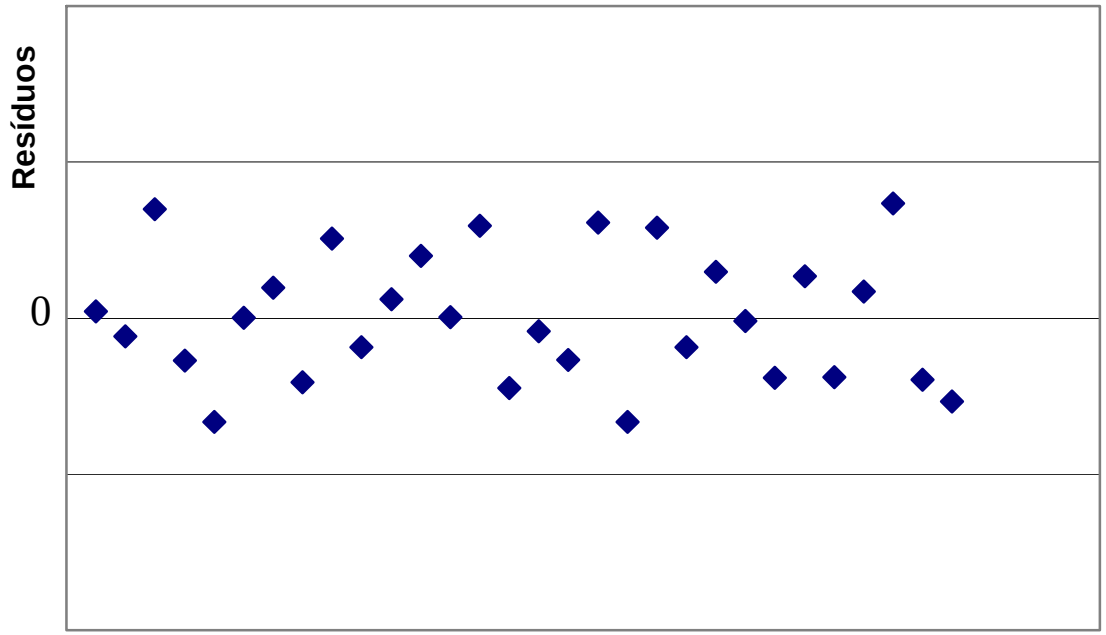
Normal Q-Q Plot



Da mesma forma, da observação do Q-Q Plot, verifica-se a presunção de normalidade pois os resíduos estão aproximadamente em linha recta.

MÉDIA NULA, VARIÂNCIA CONSTANTE E INDEPÊNDENCIA DOS ERROS

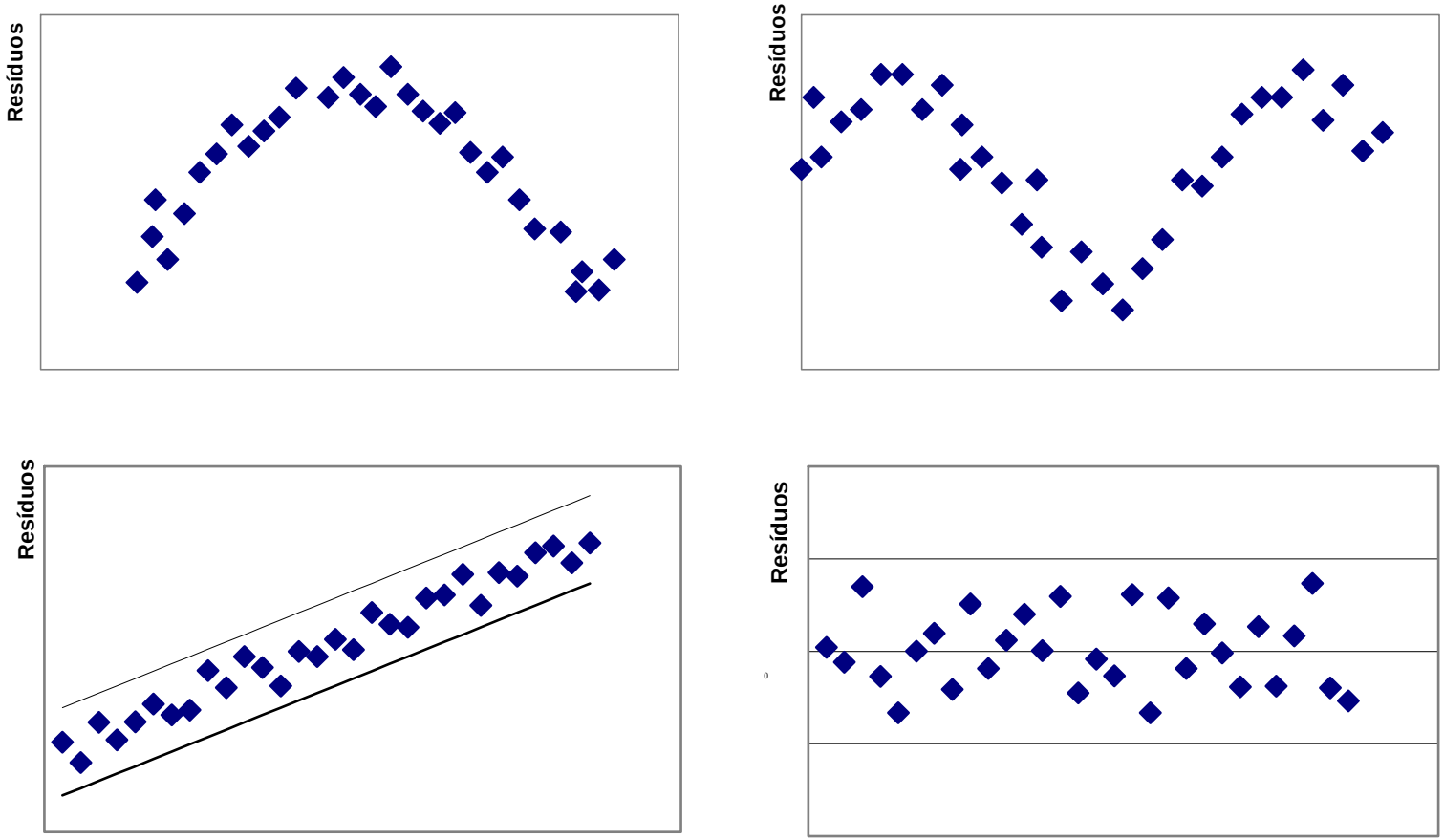
Estes pressupostos podem ser verificados graficamente, representando os resíduos em função dos valores estimados da variável dependente $\hat{y}_i$ (gráfico residual) ou em função dos valores duma das variáveis independentes x_i.



Os pontos do gráfico devem distribuir-se de forma **aleatória** em torno da recta que corresponde ao **resíduo zero,** formando uma **mancha de largura uniforme.** Dessa forma será de esperar que os erros sejam independentes, de média nula e de variância constante.

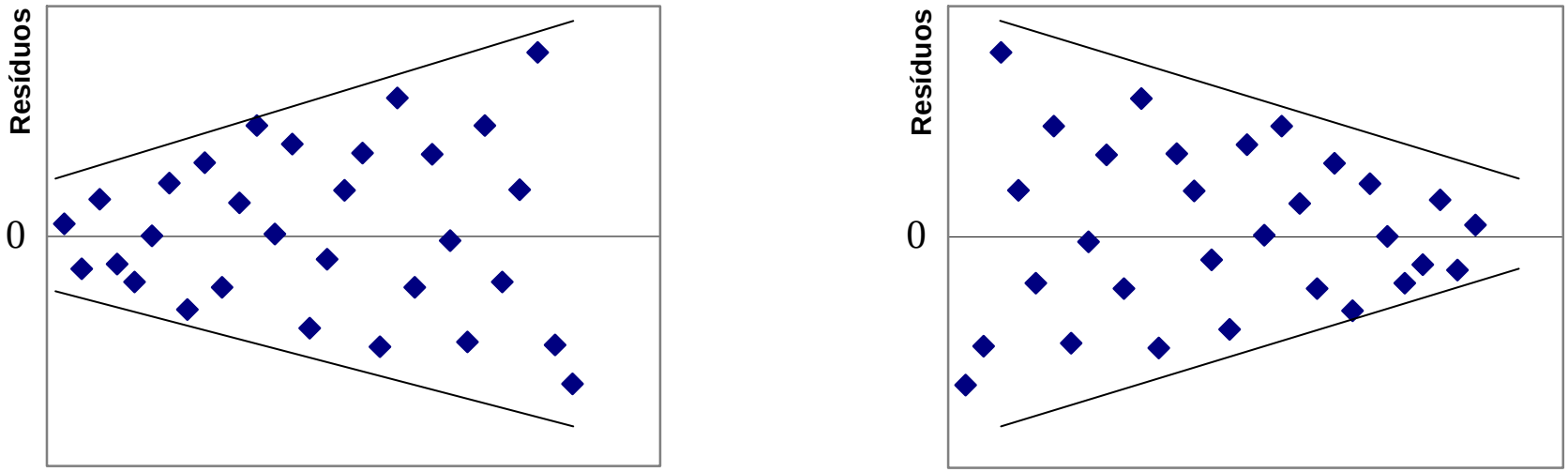
Quando os resíduos não se comportam de forma aleatória, ou seja, seguem um padrão, a condição de independência não é satisfeita.

Isto pode traduzir o facto de não existir uma relação linear entre as variáveis ou então, não constam no modelo uma ou várias variáveis independentes que influenciam significativamente a variável dependente e portanto também os erros.



Nos 3 primeiros gráficos, os resíduos apresentam comportamentos padronizados, logo não há independência. Pelo contrário, no último gráfico os resíduos parecem estar distribuídos de forma aleatória, sustentando a independência dos erros.

Se a dispersão dos resíduos aumentar ou diminuir com os valores das variáveis independentes x_i, ou com os valores estimados da variável dependente $\hat{y}_i$, deve ser posta em causa a hipótese de variâncias constante dos **E_i**'s.



No gráfico da esquerda, os resíduos apresentam um comportamento tendencialmente crescente, no da direita, o comportamento é tendencialmente decrescente, indicando que há violação da hipótese de homogeneidade da variância.

Usando um gráfico residual, as violações dos pressupostos do modelo não são sempre fáceis de detectar e podem ocorrer apesar dos gráficos parecerem bem comportados.

A análise de resíduos, usando gráficos residuais é um método subjectivo.

Nesse sentido, a verificação da independência é usualmente feita através do **teste de Durbin-Watson** à correlação entre resíduos sucessivos.

Se houver independência, a magnitude de um resíduo não influencia a magnitude do resíduo seguinte. Neste caso, a correlação entre resíduos sucessivos é nula ($\rho = 0$). As hipóteses do teste, para aferir se a relação entre dois resíduos consecutivos é estatisticamente significativa, são então:

$$\begin{array}{|ll|} \hline H_0 : \rho = 0 & \textit{existe independência} \\ H_1 : \rho \neq 0 & \textit{existe dependência} \\ \hline \end{array}$$

Estatística ***d*** de Durbin-Watson:

$$d = \frac{\sum_{i=1}^{n-1}\left(d_{i+1} - d_i\right)^2}{\sum_{i=1}^{n} d_i^{\,2}}.$$

Tomada de decisão:

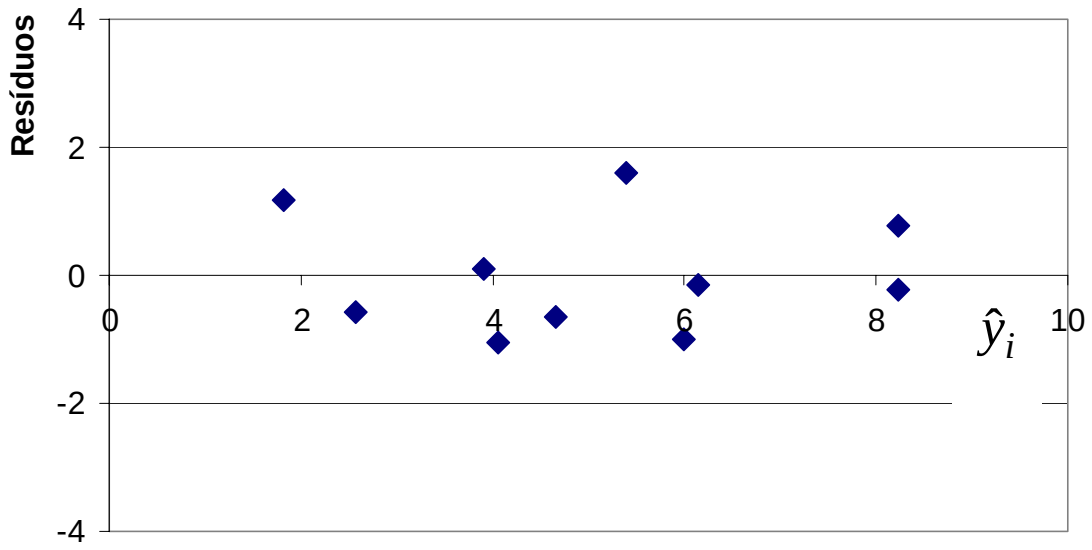
Compara-se o valor obtido para a estatística ***d*** com os valores críticos da tabela de Durbin-Watson, d_L e d_U, e toma-se a decisão recorrendo à seguinte tabela:

d	$[0, d_L[$	$[d_L, d_U[$	$[d_U, 4-d_U[$	$[4-d_U, 4-d_L[$	$[4-d_L, 4[$
Decisão	Rejeitar H_0 **Dependência**	**Nada se pode concluir**	Não rejeitar H_0 **Independência**	**Nada se pode concluir**	Rejeitar H_0 **Dependência**

Só quando $d \in [d_U, 4-d_U[$, se pode concluir que os diferentes valores de E_i são independentes.

Exemplo:

Construímos o gráfico residual relativo ao exemplo em estudo.



A análise gráfica dos resíduos, dá indicação de que os resíduos parecem distribuir-se aleatoriamente à volta da recta x=0, com dispersão constante, sugerindo que não há violações sérias dos pressupostos de homocedasticidade, média nula e de independência dos erros.

Para verificar o pressuposto de independência vamos, também, utilizar o teste de Durbin-Watson.

Com os dados:

Vendedor	d_i	d_i^2	d_{i+1} - d_i	$(d_{i+1} - d_i)^2$
1	-1,05932	1,122159	-0,91526	0,837700868
2	-0,99153	0,983132	-0,50847	0,258541741
3	-0,65254	0,425808	1,83051	3,35076686
4	-0,5678	0,322397	-2,23729	5,005466544
5	-0,22881	0,052354	0,06779	0,004595484
6	-0,14407	0,020756	0,76272	0,581741798
7	0,09322	0,00869	-0,33899	0,11491422
8	0,77119	0,594734	2,16949	4,70668686
9	1,17797	1,387613	-1,50847	2,275481741
10	1,60169	2,565411		
Soma		7,48305		17,13589612

Obtém-se:

$$d = \frac{17{,}13589612}{7{,}48305} = 2{,}28996$$

Com n = 10, k = 2 e α = 0.05, os valores críticos da tabela de Durbin-Watson são:

$$d_L = 0.7 \text{ e } \quad d_U = 1.64$$

e,

$$[d_U, 4 - d_U[= [1.64,\ 4\text{-}1.64[= [1.64,\ 2.36[$$

Uma vez que d=2.28996 $\in$ [1.64, 2.36[, não é rejeitada a hipótese de independência. Podemos pois admitir que os erros são independentes, ou seja, que se verifica o pressuposto da independência, o que vai de encontro ao que verificamos graficamente.