# Modelos de Regressão Linear Simples - Análise de Resíduos

Erica Castilho Rodrigues

27 de Setembro de 2016

- O modelo de regressão linear é dado por

- O modelo de regressão linear é dado por

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

onde

$$\epsilon_i$$

- O modelo de regressão linear é dado por

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

onde

$$\epsilon_i \sim^{iid} N(0, \sigma^2).$$

- O erro $\epsilon_i$ é estimado pelo resíduo $e_i$

$$e_i = \hat{Y}_i - Y_i \, .$$

- Representa a quantidade da varilibilidade que $Y$ que o modelo ajustado não consegue explicar.

- Os resíduos contém informação sobre o motivo do modelo não ter se ajustado bem aos dados.
- Conseguem indicar se uma ou mais suposições do modelo foram violadas.
- Principais problemas detectados através da análise dos resíduos:
  - Não-linearidade da relação entre X e Y ;
  - Não normalidade dos erros;
  - Variância não-constante dos erros (heterocedasticidade);
  - Correlação entre os erros;
  - Presença de *outliers* ou observações atípicas;
  - O modelo foi mal especificado.

- Vejamos as suposições do modelo com mais detalhes.

## Independência

- O modelo supõe que os erros são independentes entre si.
- Logo os erros devem ser não correlacionados.
- Em algumas situações práticas essa suposição pode não ser verdadeira.
- Exemplos:
  - medidas repitidas $\Rightarrow$ coleta-se a medida em um mesmo indivíduo em diferentes instantes de tempo;
  - série temporal $\Rightarrow$ os dados possuem estrutura temporal que não é captada pelo modelo;
  - dados hierárquicos $\Rightarrow$ indivíduos agrupados, por exemplo, alunos em uma escola.

## Identicamente Distribuídos

- Uma das suposições é que os erros são identicamente distribuídos com distribuição $N(0, \sigma^2)$.
- Ou seja, todos erros $\epsilon_i$ foram gerados de uma mesma normal, com mesma média e variância.

## Linearidade

- O modelo supõe que $X$ e $Y$ possuem uma relação linear.
- Essa relação pode não ser linear e mesmo assim $X$ e $Y$ podem estar correlacionadas.
- Outros tipos de modelos, como Splines, polinômios, podem ser usados.

## Gráficos para análise de resíduos

- Gráfico de Probabilidade Normal dos resíduos;
- Gráfico dos resíduos versus valores de $\hat{Y}$;
- Gráfico dos resíduos versus valores de $X$ (incluída no modelo);
- Gráfico dos resíduos versus outras Xs (não incluídas no modelo);
- Gráfico dos resíduos versus tempo ou ordem de coleta dos dados.

# Verificação de Não-Normalidade dos Erros

- Assumimos que os erros $\epsilon_i \sim N(0, \sigma^2)$ para $i = 1, \ldots, n$.
- Desvios da normalidade afetam:
  - os intervalos de confiança;
  - testes $t$ e $F$.
- Usamos os resíduos como estimativa do erro para verificar a suposição.

- Para testar normalidade podemo usar:
  - Histograma: deve ser simétrico em torno de zero;
  - Gráfico de Probabilidade Normal: verifica visualmente se os dados seguem uma normal;
  - Testes de normalidade (Shapiro-Wilk, Anderson Darling).
    - A hipótese nula é de que os dados são normais e deverá ser rejeitada se o p-valor é pequeno.
    - Vamos usar aqui mais o Gráfico de Probabilidade Normal.

# Gráficos de Probabilidade Normal

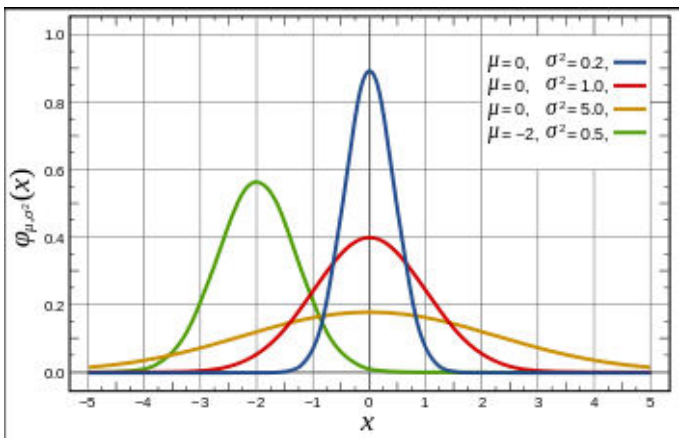
- Seja $X$ uma variável aleatória $N(\mu, \sigma^2)$.

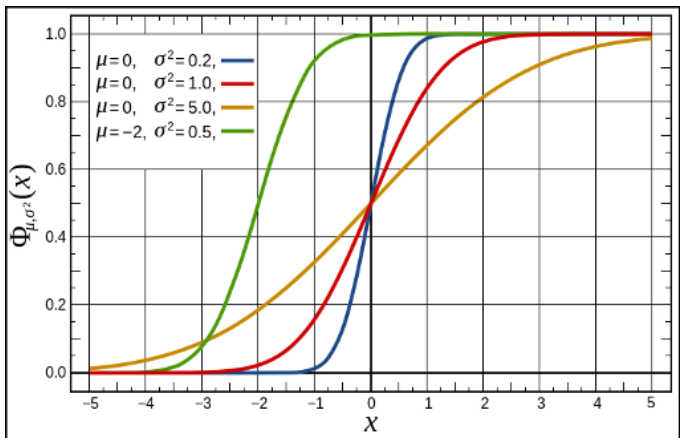
- A função densidade de $X$ é dada por

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- A função de distribuição acumulada é dada por

$$F_X(x) = P(X < x) = \int_{-\infty}^{x} f_X(t)dt$$



- O histograma dá uma idéia da distribuição dos dados:
  - apenas se amostra é grande.

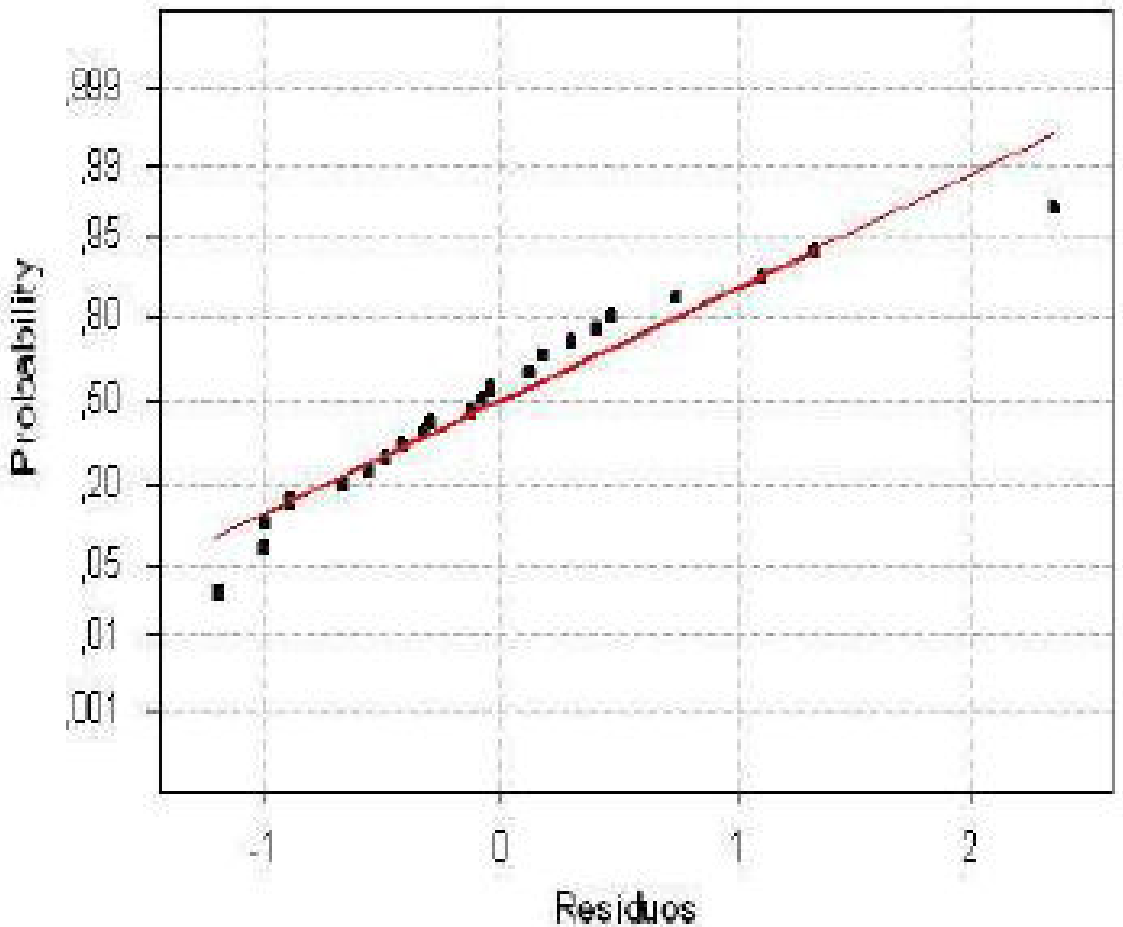
## Gráfico de Probabilidade

- É o gráfico de $F_X(x)$ em uma escala especial.
- Determina se os dados obedecem uma distribuição hipotética.
- Baseado no exame visual dos dados.
- Os pontos plotados no gráfico são:
  - $e_i$ e ordem do percentil de $e_i$, ou seja,

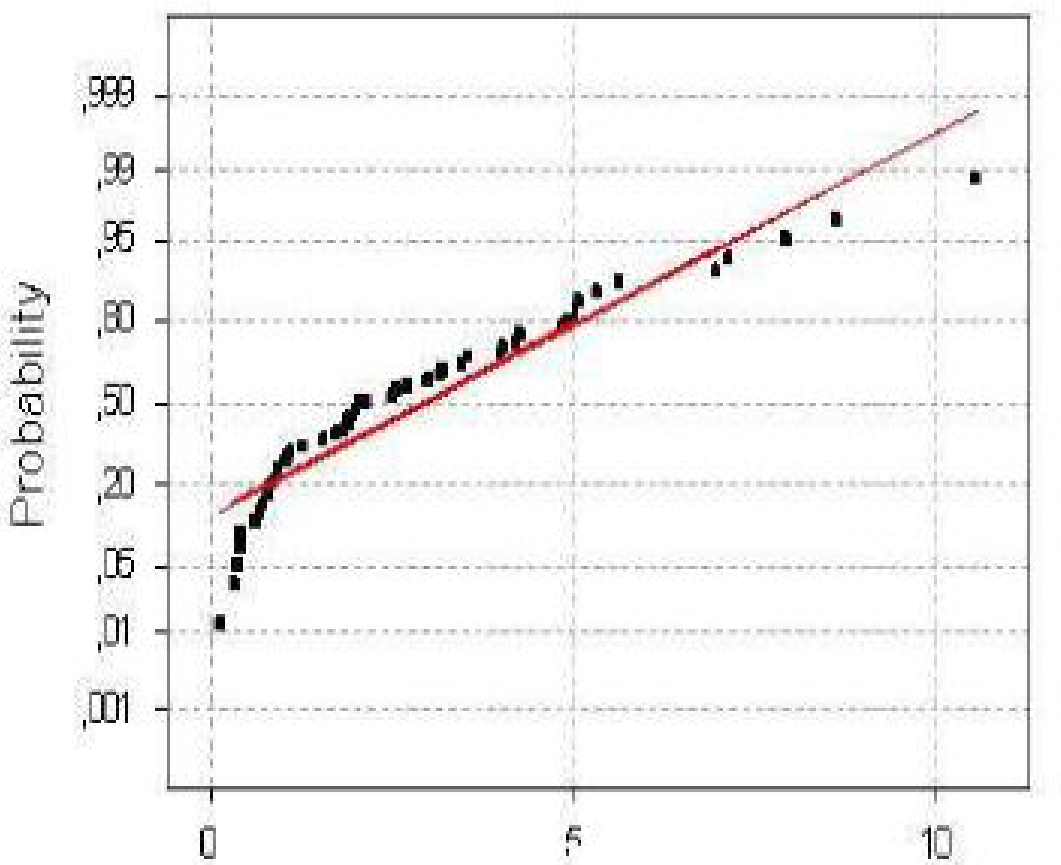
$$(e_i;\ \%\text{ de } e_i\text{'s } < e_i)\,.$$

- Se os pontos caem em torno de uma reta a distribuição é adequada.

- Distribuição normal



- Distribuição assimétrica



## Exemplo:

- Dez observações da corrente em um fio foram coletadas.
- Queremos verificar se seguem uma distribuição normal.

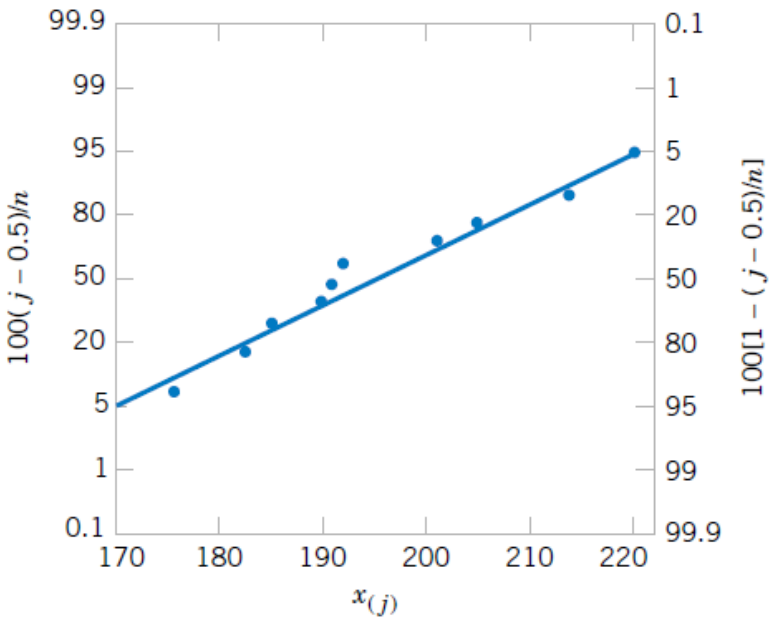



Figure 6-19 Normal probability plot for battery life.

## Exemplo:

- Dez observações da corrente em um fio foram coletadas.
- Queremos verificar se seguem uma distribuição normal.

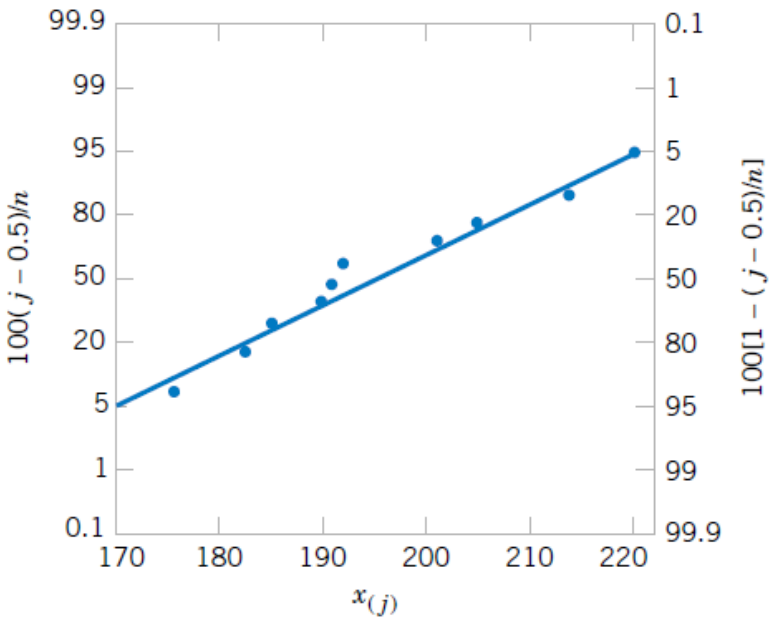
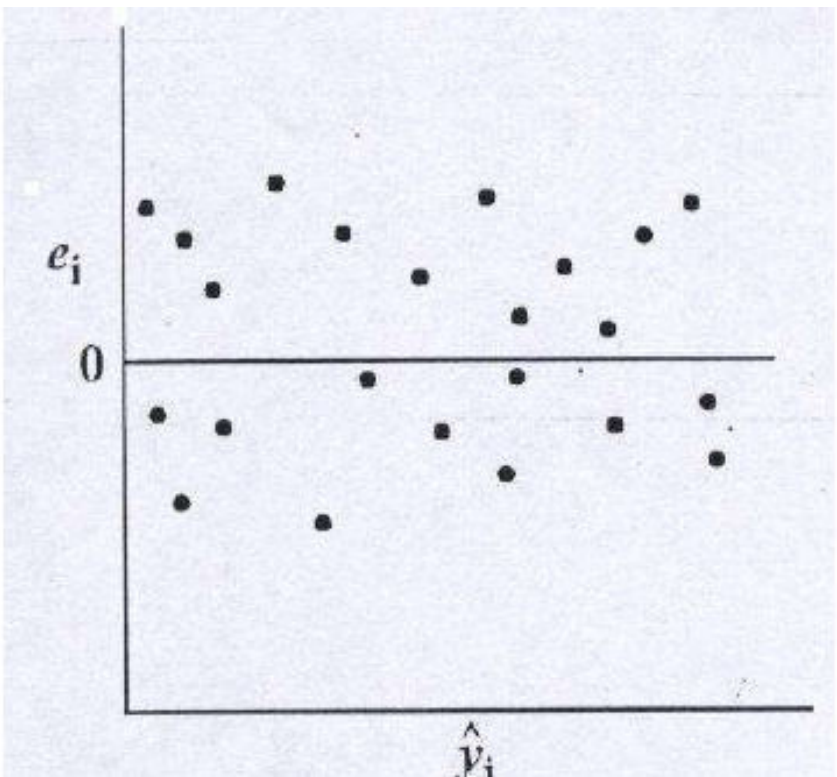



Figure 6-19 Normal probability plot for battery life.

- Os pontos caem aproximadamente em torno da linha.
- Isso indica que os dados têm distribuição normal.

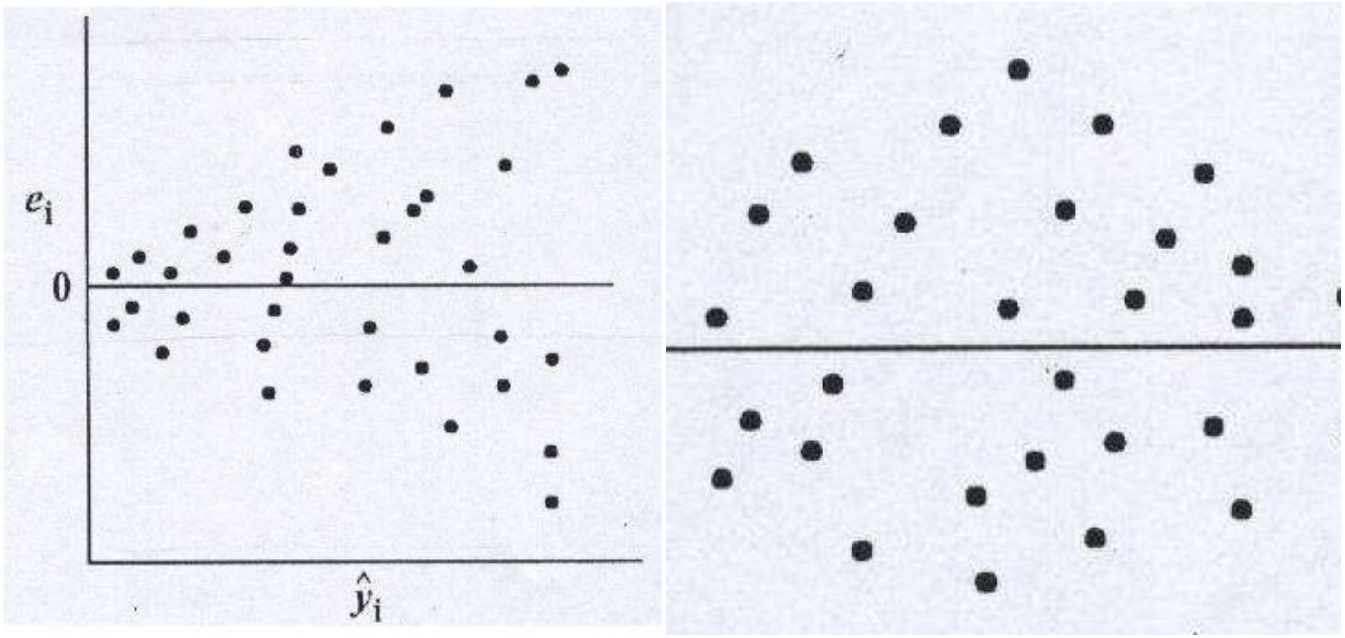
# Gráfico dos Resíduos $e_i$ contra Valores Ajustados $\hat{Y}_i$

- Aparência desejada:
  - nuvem de pontos aleatória e homogênea em torno do eixo horizontal $Y = 0$.



Útil para detectar as seguintes inadequções do modelo:

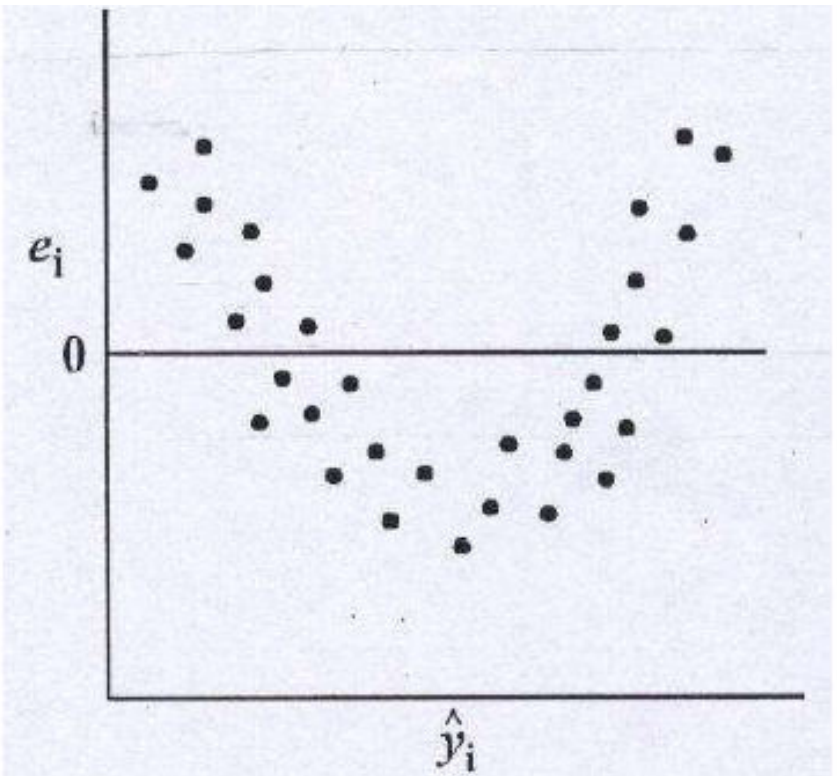
- A variância do erro não é constante.
  - Solução: fazer transformação em $Y$ ou usar Mínimos Quadrados Ponderados.



**A homocedasticidade é provavelmente violada se...**

- Se os resíduos aumentam ou diminuem com os valores ajustados.
- Se os pontos formam uma curva ao redor de zero e não estão dispostos aleatoriamente.
- Poucos pontos no gráfico ficam muito distantes dos demais.

- A equação de regressão não é linear.
  - Solução: transformações em $Y$ e/ou $X$; inclusão do termo quadrático de X.



## Gráfico dos Resíduos contra a Variável Explicativa

- Na Regressão Linear Simples, tem o mesmo papel do gráfico $e_i$ vs $\hat{Y}_i$.
- Em Regressão Múltipla, pode ser usado para verificar a necessidade de se incluir variáveis.
- Nesse último caso, é feito o gráfico dos resíduos vs variáveis não incluídas no modelo.
- Se houver algum padrão, significa que a variável deve ser incluída.

# Gráfico dos Resíduos contra o Tempo ou Ordem de Coleta

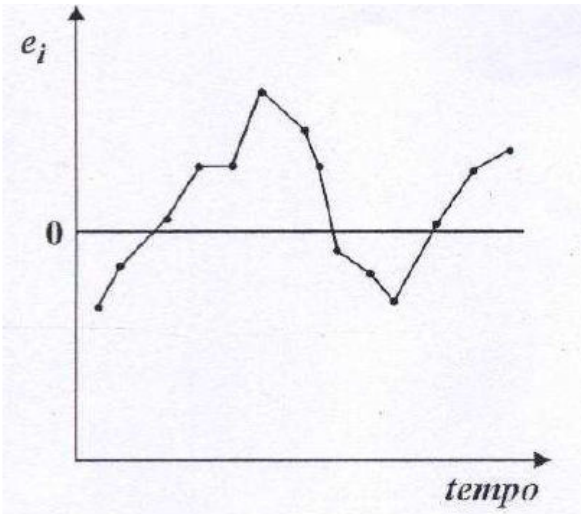
- Os erros devem ser independentes entre si.
- Esse gráfico verifica apenas se eles estão correlacionados no tempo.
- Só pode ser usado caso os dados sejam coletados sequencialmente.
- Os erros são plotados na ordem em que foi feita a coleta.
- A presença de algum padrão indica correlação entre eles.
- A existência de correlação temporal pode ser pode ser consequência da:
  - não inclusão que uma variável explicativa relacionada ao tempo.

## Autocorrelação

É a correlação entre o erro no tempo $t$ e os erros dos tempos anteriores $(t-1, t-2, \ldots)$.

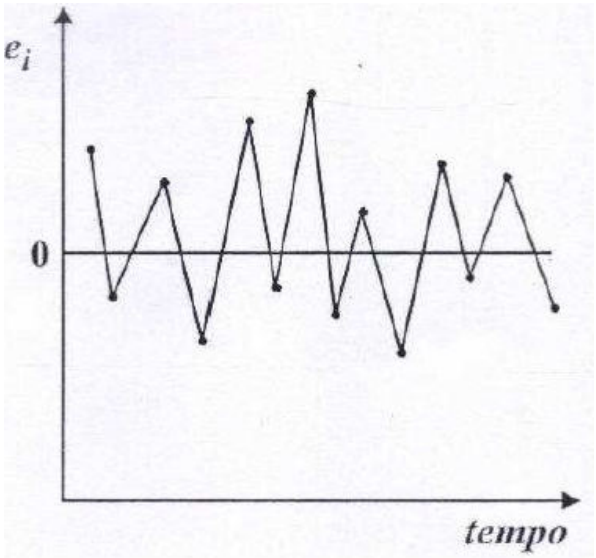
### Autocorrelação Positiva

- Se um erro está acima de zero, o próximo tende a estar também.




### Autocorrelação Negativa

- Se um erro está acima de zero, o próximo tende a estar abaixo.



## Consequências das correlações entre os erros

- Os estimadores de Mínimos Quadrados deixam de ser bons estimadores.
- Os intervalos de confiança e testes não são mais apropriados.

## Teste de Durbin-Watson

- Testa se existe dependência sequencial entre os erros.
- Verifica se cada erro está correlacionado com o anterior.
- A estatística de teste é dada por:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}.$$

- Essa estatística está sempre no intervalo $[0, 4]$.
- É distribuída simetricamente em torno de 2.
- Se os erros tem correlação positiva $\Rightarrow d \approx 0$.
- Se os erros tem correlação negativa $\Rightarrow d \approx 4$ ou $4 - d \approx 0$.

- O teste é feito usando a seguinte tabela.

| | 1% | | 2.5% | | 5% | |
|---|---|---|---|---|---|---|
| n | $d_L$ | $d_U$ | $d_L$ | $d_U$ | $d_L$ | $d_U$ |
| 15 | 0.81 | 1.07 | 0.95 | 1.23 | 1.08 | 1.36 |
| 20 | 0.95 | 1.15 | 1.08 | 1.28 | 1.20 | 1.41 |
| 25 | 1.05 | 1.21 | 1.18 | 1.34 | 1.29 | 1.45 |
| 30 | 1.13 | 1.26 | 1.25 | 1.38 | 1.35 | 1.49 |
| 40 | 1.25 | 1.34 | 1.35 | 1.45 | 1.44 | 1.54 |
| 50 | 1.32 | 1.40 | 1.42 | 1.50 | 1.50 | 1.59 |
| 70 | 1.43 | 1.49 | 1.51 | 1.57 | 1.58 | 1.64 |
| 100 | 1.52 | 1.56 | 1.59 | 1.63 | 1.65 | 1.69 |
| 150 | 1.61 | 1.64 | — | — | 1.72 | 1.75 |
| 200 | 1.66 | 1.68 | — | — | 1.76 | 1.78 |

- O teste é feito da seguinta maneira:

Para $d < 2$ olhamos para $d$

$$\text{Para d<2 : } \begin{cases} d < d_L & \text{possível correlação serial positiva} \\ d > d_U & \text{nenhuma indicação de correlação serial} \\ d_L < d < d_u & \text{teste inconclusivo} \end{cases}$$

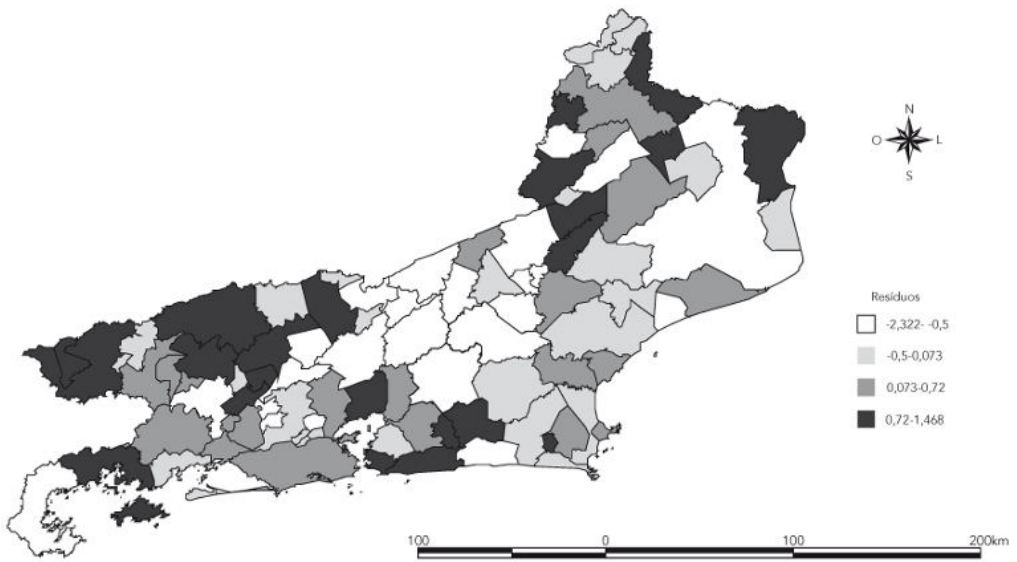
Para $d > 2$ olhamos para $4 - d$

$$\text{Para d>2 : } \begin{cases} 4 - d < d_L & \text{possível correlação serial negativa} \\ 4 - d > d_U & \text{nenhuma indicação de correlação serial} \\ d_L < 4 - d < d_u & \text{teste inconclusivo} \end{cases}$$

- Os erros podem também estar espacialmente correlacionados.
- Isso acontece se omitimos uma variável com dependência espacial.
- Podemos fazer o mapa e verificar se existe padrão espacial.

Figura 2

Mapa dos resíduos do modelo de regressão linear multivariada.



## Resumindo...

- Os erros são assumidos não correlacionados e com variância constante.
- Usamos os resíduos (estimativas do erro) para verificar essas suposições.

| Gráfico dos Resíduos | Suposições Avaliadas |
|---|---|
| $e_i$ vs $\hat{Y}_i$<br>$e_i$ vs $X_i$ | Variância Constante<br>Linearidade |
| $e_i$ vs Variáveis não incluídas | Suficiência das variáveis incluídas. |
| Probabilidade Normal | Normalidade |
| $e_i$ vs tempo de coleta | Ausência se autocorrelação temporal. |

## Exemplo

- Considere os dados do consumo de gás.
- Lembre-se que

$$Y = \{\text{Temperatura Atmosférica do Mês}\}$$

$$X = \{\text{Consumo Mensal de Gás Residencial}\}$$

- O modelo de regressão ajustado é dado por

## Exemplo

- Considere os dados do consumo de gás.
- Lembre-se que

$$Y = \{\text{Temperatura Atmosférica do Mês}\}$$

$$X = \{\text{Consumo Mensal de Gás Residencial}\}$$

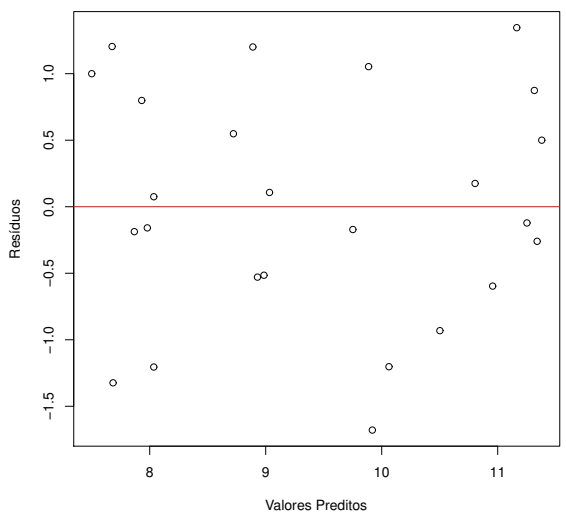
- O modelo de regressão ajustado é dado por

$$Y_i = 13.6230 - 0.0798 X_i + \epsilon_i$$

onde $\epsilon_i \sim^{iid} N(0, \sigma^2)$.

## Exemplo (continuação)

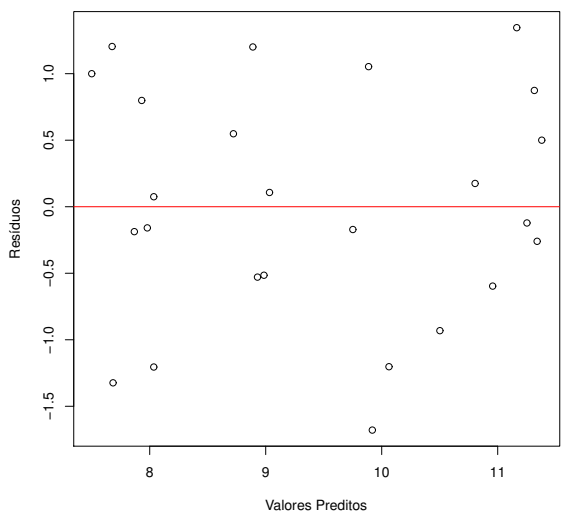
- A figura abaixo mostra o gráfio dos resíduos vs valores ajustados.



## Exemplo (continuação)

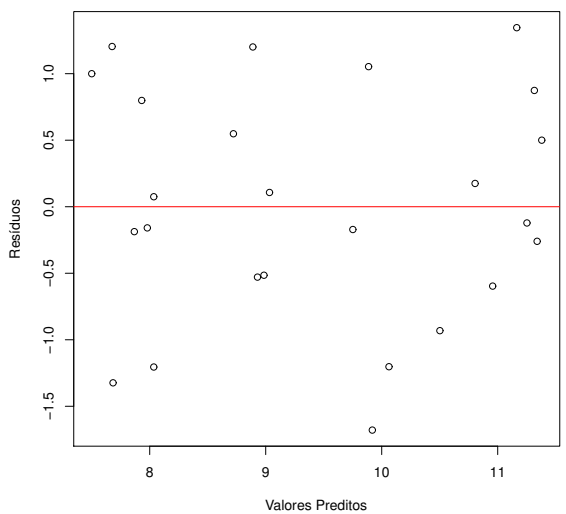
- A figura abaixo mostra o gráfio dos resíduos vs valores ajustados.

Conclusões:



## Exemplo (continuação)

- A figura abaixo mostra o gráfio dos resíduos vs valores ajustados.

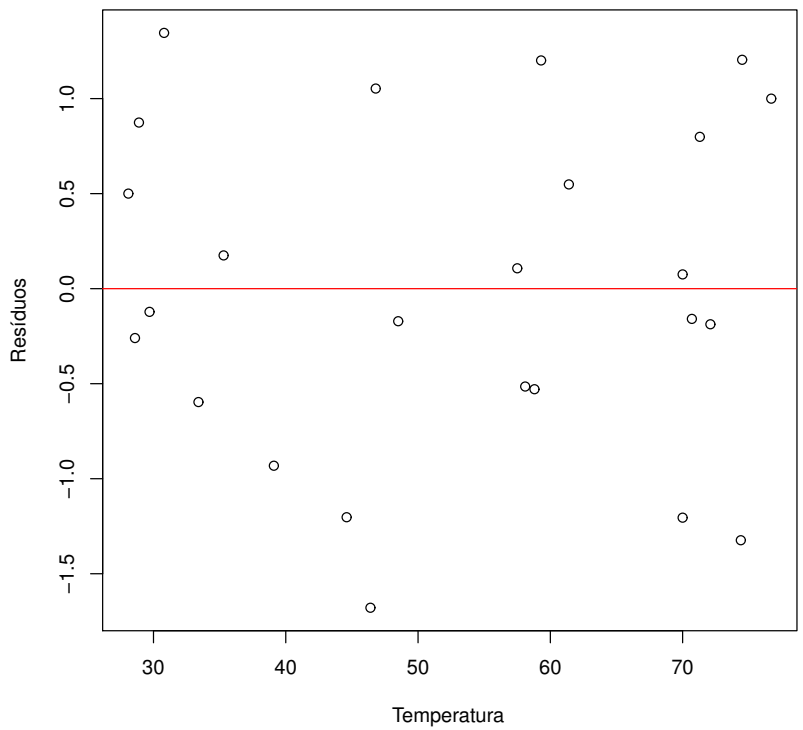



Conclusões:

- Os resíduos se distribuem aleatoriamente em torno de zero.
- Não se observa nenhum padrão.
- Isso indica que:
  - a variância é constante;
  - a relação entre as variáveis é linear.

## Exemplo (continuação)

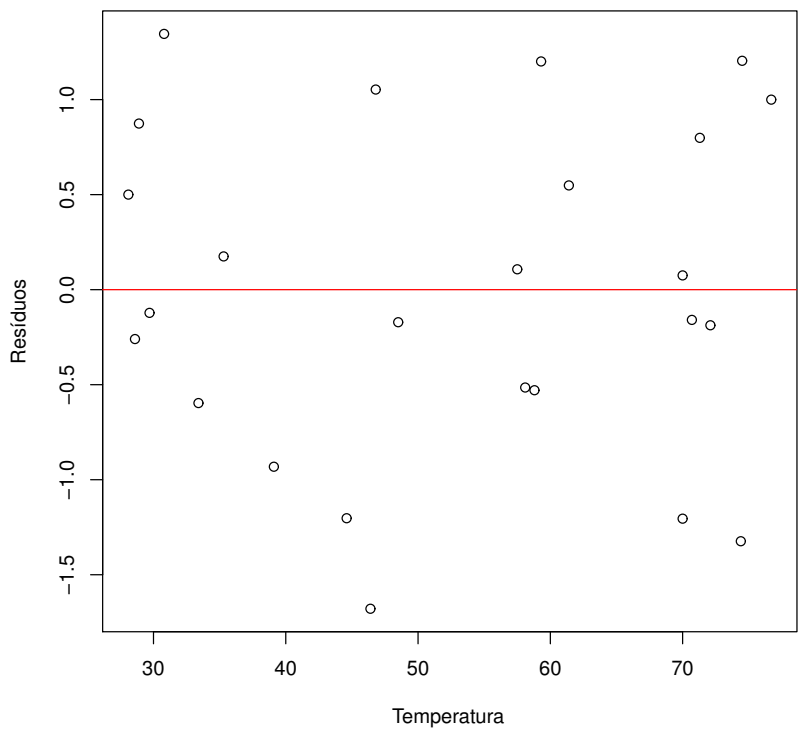
- A figura abaixo mostra o gráfico dos resíduos vs variável preditora.




- Conclusões:

## Exemplo (continuação)

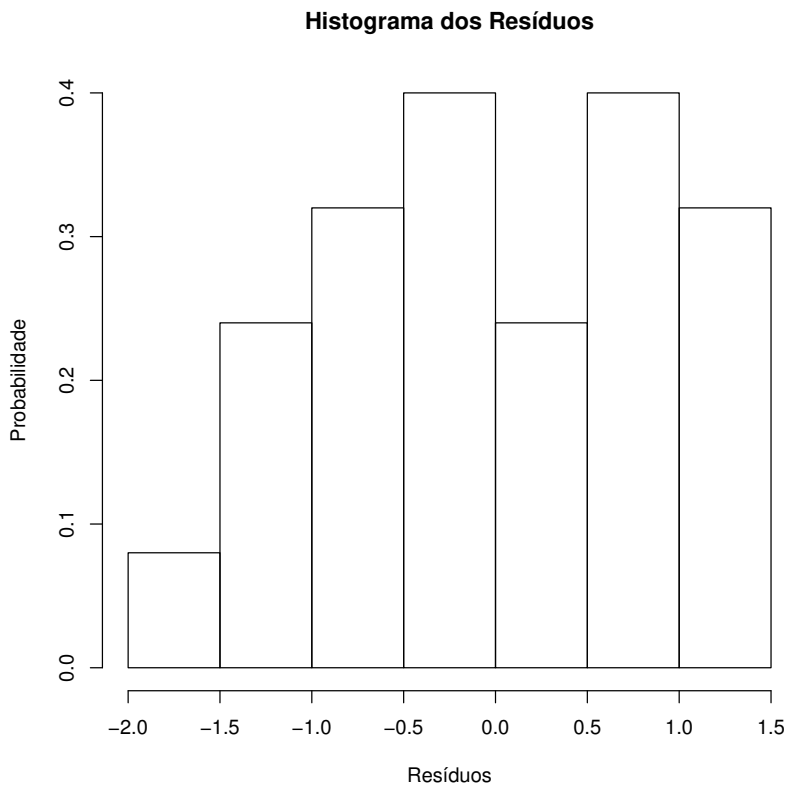
- A figura abaixo mostra o gráfico dos resíduos vs variável preditora.

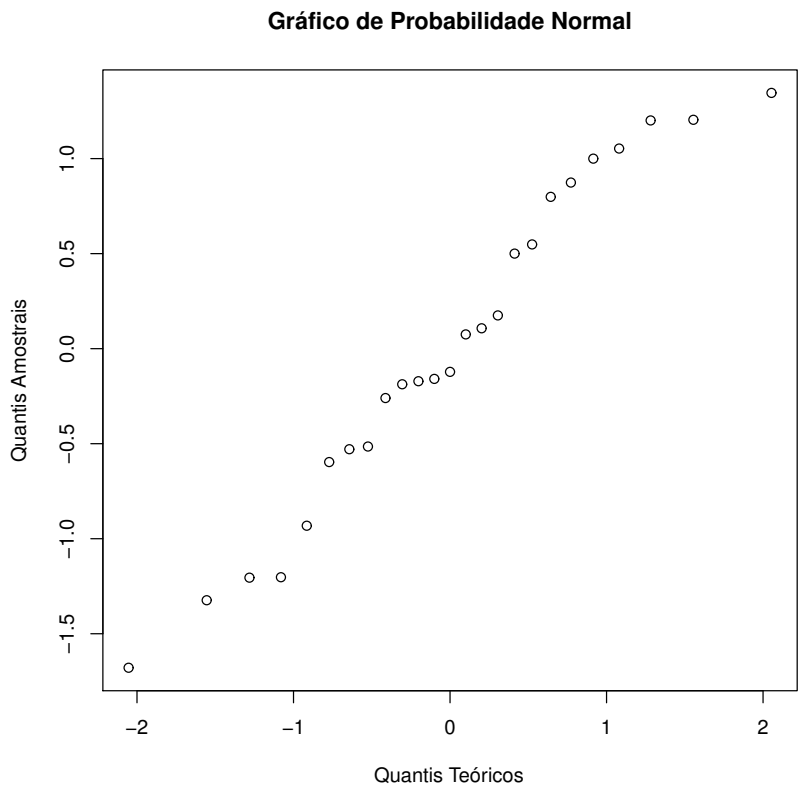



- Conclusões: são as mesmas do gráfico anterior.

# Exemplo (continuação)

- As figuras abaixo mostram o histogram e gráfico de probabilidade normal dos resíduos.

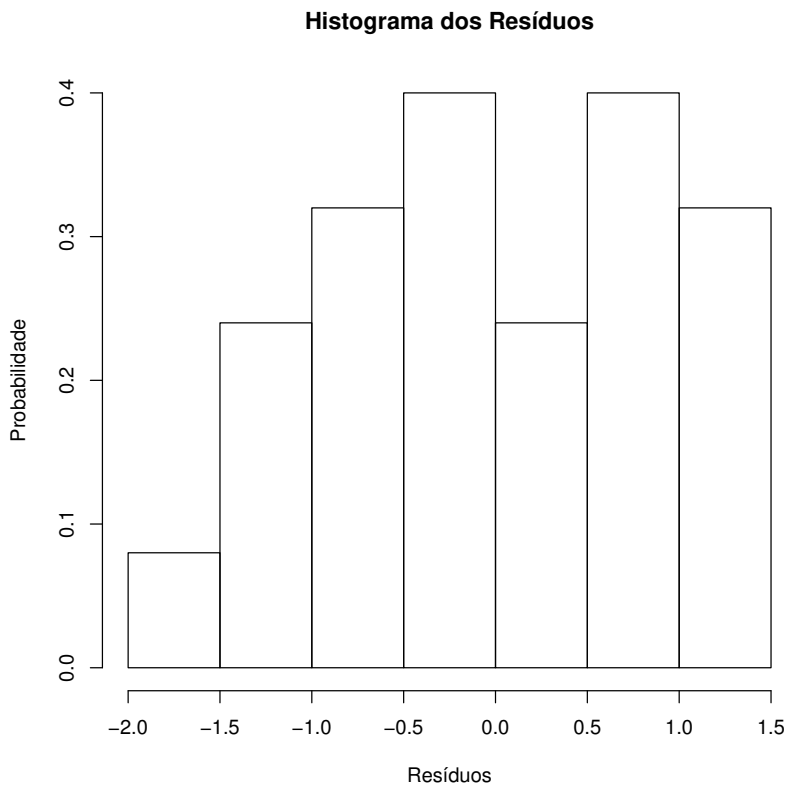


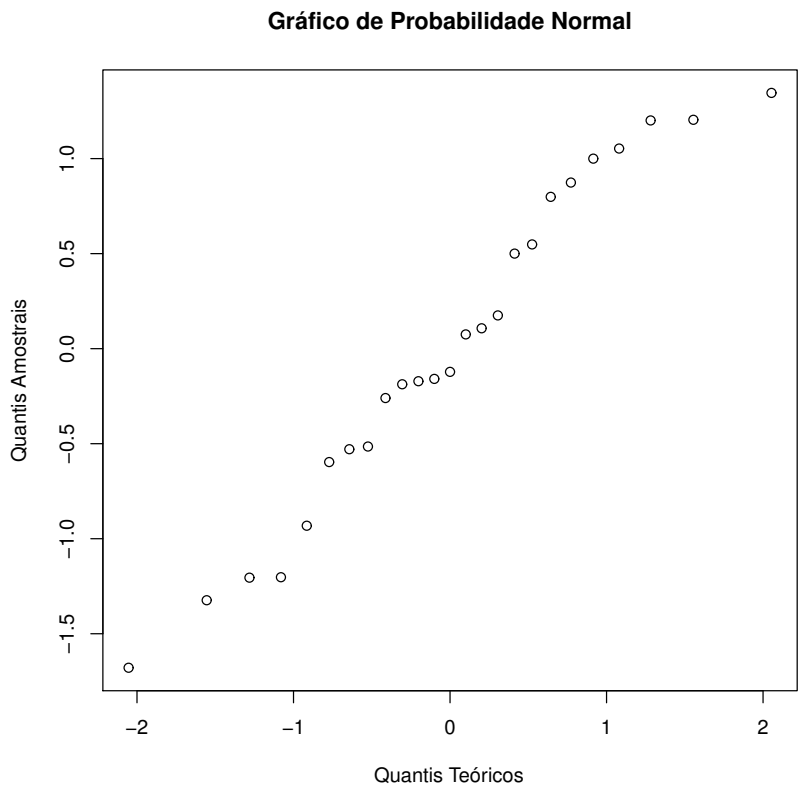




- Conclusão:

# Exemplo (continuação)

- As figuras abaixo mostram o histogram e gráfico de probabilidade normal dos resíduos.




- Conclusão: os resíduos parecem seguir uma distribuição normal

## Exemplo (continuação)

- Gráfico de resíduos em função do tempo.
- Podemos fazer esse gráfico para esse problema?

## Exemplo (continuação)

- Gráfico de resíduos em função do tempo.
- Podemos fazer esse gráfico para esse problema?
- Não, pois não sabemos a ordem de coleta.

# *Outliers*

## *Outliers*

Observação numericamente distante do restante dos dados.

## Como idenficar *Outliers*

- A i-ésima observação será *outlier* se:
  - $e_i$ for maior que do que dois desvios padrões.
- Lembre que o desvio padrão é estimado pela $\sqrt{S^2}$.
- Pontos muito distantes nos gráficos são *outliers*.

## Porque aparecem?

- Erros de digitação.
- Assimetria da distribuição.
- Aleatoridade.

## O que fazer?

- Eliminar?
- Corrigir?
- Analisá-los?
- Usar um modelo robusto a *outliers*?

- Podemos medir a influência dessas observações atípicas.
- Uma das possibilidades: *Cook's distance*

## Cook's distance

- É uma medida de distância calulada para cada ponto da base de dados.
- É dada por

$$D_i = \frac{\sum_{i=1}^{n} (\hat{y}_j - \hat{y}(i)_j)^2}{(p+1)S^2}$$

onde

- $\hat{y}_j$ é o valor ajustado usando todos os dados;
- $\hat{y}(i)_j$ valor ajustado removendo a i-ésima observação;
- $S^2$ é estimativa de $\sigma^2$;
- $p$ é o número de variáveis no modelo.

- Mede o quanto que o modelo muda ao descartarmos $y_i$.
- Se $y_i$ é um ponto muito influente:
  - o modelo muda muito;
  - os valores $\hat{y}_j$ ficam muito distantes de $\hat{y}(i)_j$.
- Se $y_i$ não é um ponto muito influente:
  - o modelo não muda muito;
  - os valores $\hat{y}_j$ ficam muito próximos de $\hat{y}(i)_j$.
- Dizemos que a observação $i$ é *outlier* se

$$D_i > \frac{4}{n-(p+1)} \text{ no caso univariado } D_i > \frac{4}{n}\,.$$

- Esse método apenas identifica pontos que são *outliers*.
- Não devemos eliminá-los imediatamente.
- A eliminação de dados é perigosa.
- Irá melhorar o asjute do modelo.
- Porém, podemos estar jogando fora informação importante dos dados.
- É necessário verificar se são erros, de fato.
- Entraremos em mais detalhes sobre isso quando estudarmos os modelos de regressão múltipla.