Tópicos em Regressão Linear – Aula 08

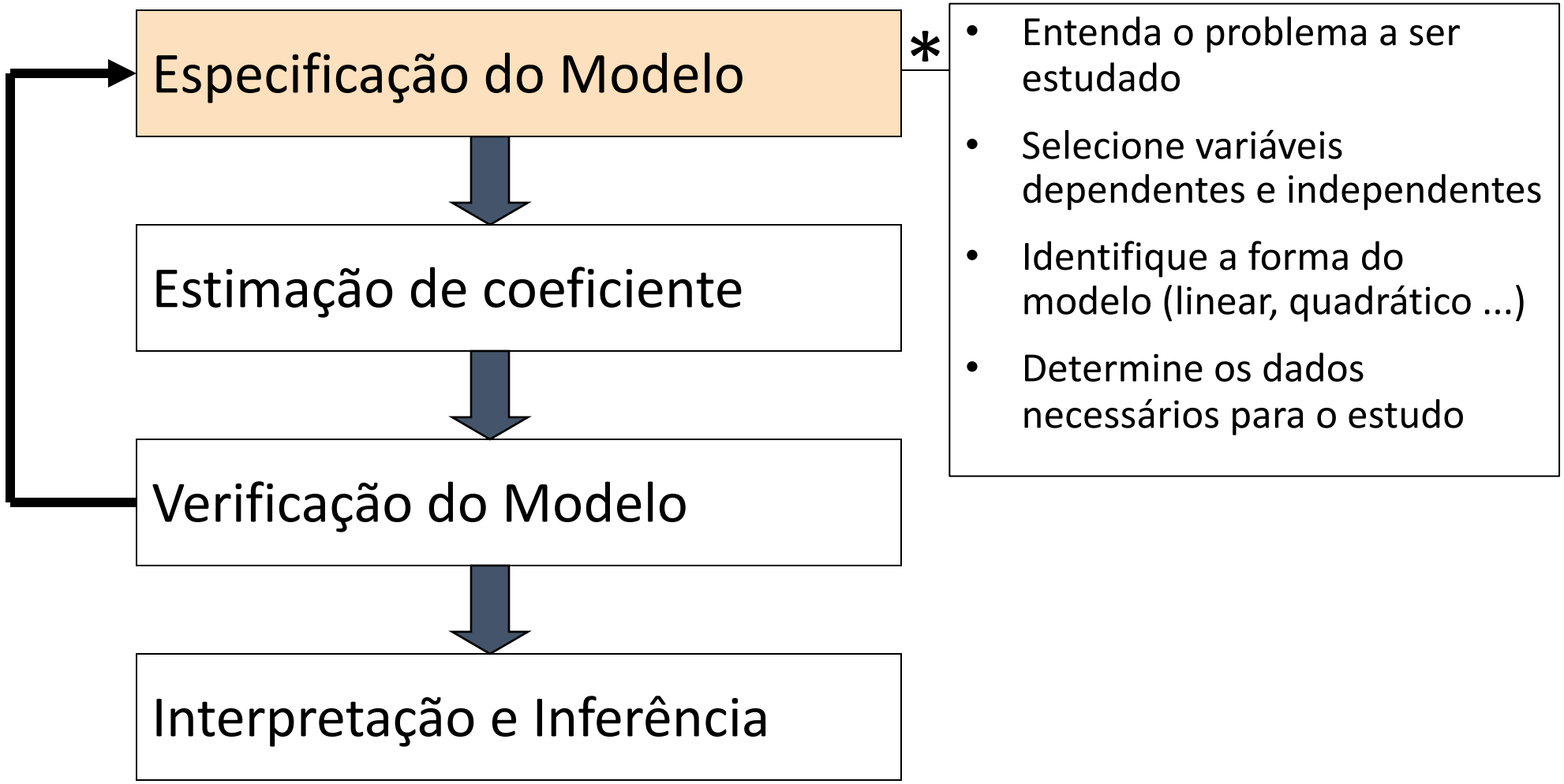
Statistics for Business and Economics 7 edição, by Paul Newbold , William Carlson , Betty Thorne (cap. Additional Topics in Regression Analysis)

Bussab e Morettin (Cap 16 Regress.o Linear Simples)

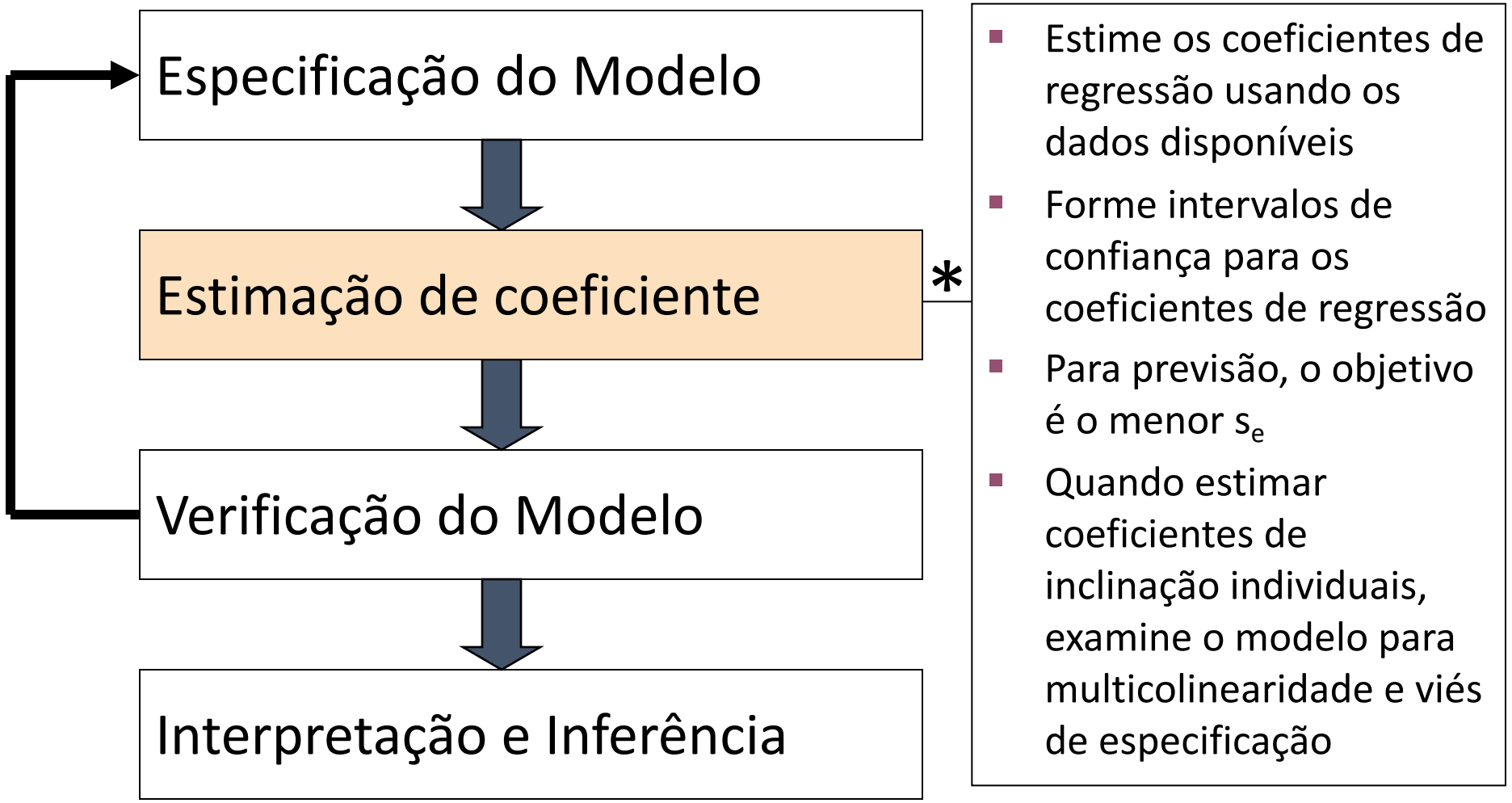
Statistics for Economics, Accounting and Business Studies, capeitulo 8, Barrow (Multiple Regression)

Marislei Nishijima

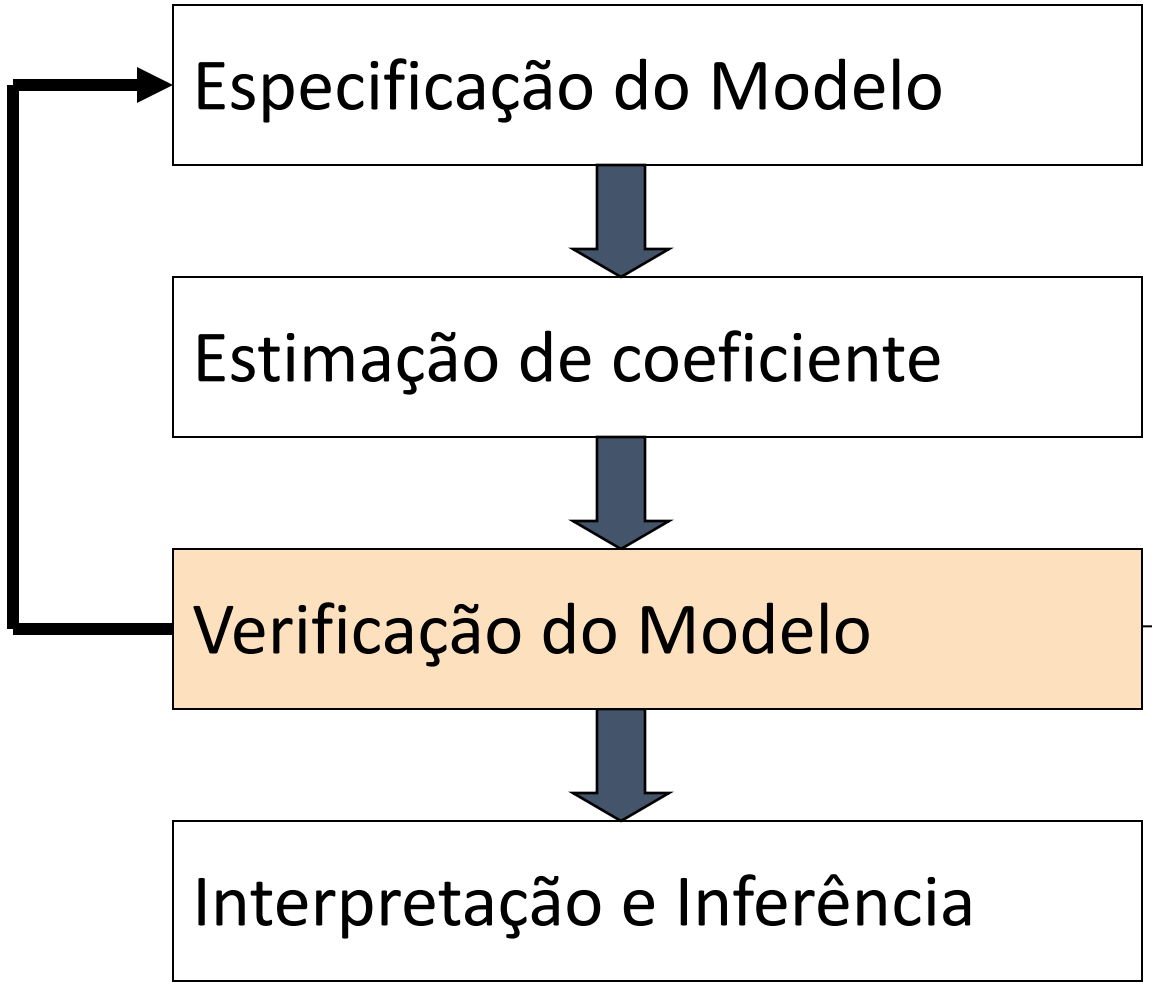
Etapas de construção do modelo



Etapas de construção do modelo



Etapas de construção do modelo



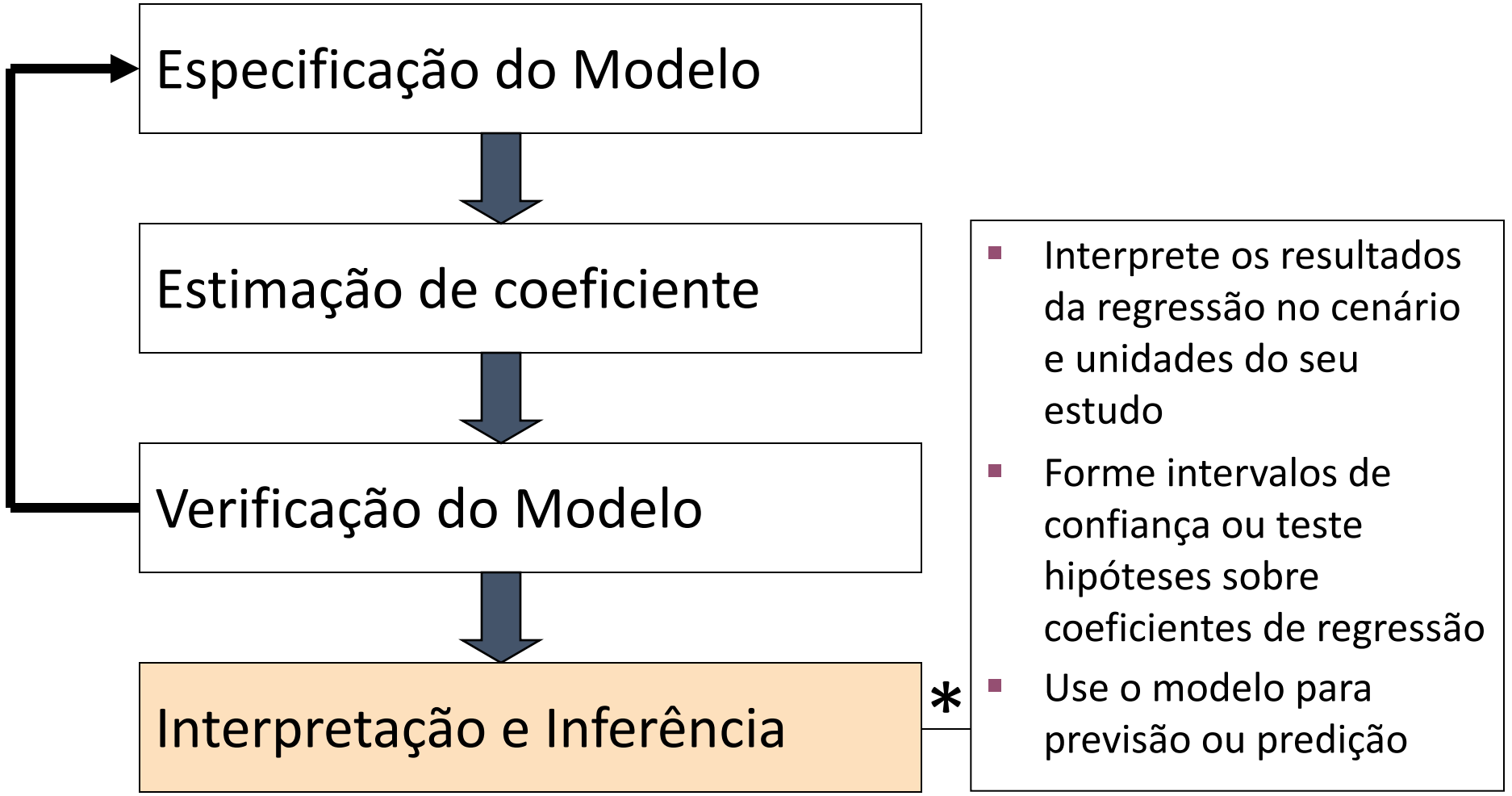
*

(cont.)

- Avalie logicamente os resultados da regressão à luz do modelo (ou seja, os sinais dos coeficientes estão corretos?)
- Algum coeficiente é tendencioso ou ilógico?
- Avalie as suposições de regressão (ou seja, os resíduos são aleatórios e independentes?)
- Se houver suspeita de problemas, retorne às especificações do modelo e ajuste o modelo

Etapas de construção do modelo

(cont.)



Modelos de variáveis dummies (mais de 2 níveis)

- Variável Dummy variables pode ser usado em situações em que a variável categórica de interesse tem mais de duas categorias

Variáveis dummies também podem ser úteis em experimentos

- Experimento é usado para identificar possíveis causas de variação no valor da variável dependente
- Os resultados Y são medidos em combinações específicas de níveis para tratamento e variáveis de bloqueio
- O objetivo é determinar como os diferentes tratamentos influenciam o resultado Y

Modelos de variáveis dummies (mais de 2 níveis)

- Considere a variável categórica com K níveis
- O número de variáveis dummies é **um a menos do que o número de níveis, K – 1**
- Exemplo:

  y = preço da casa; x_1 = metragem quadrada

- Se o estilo da casa é relevante:

  Estilo = rancho, dois andares, condomínio

Três níveis, então duas variáveis dummies são necessárias

Modelos de variáveis dummies (mais de 2 níveis)

(cont.)

- Exemplo: Seja o "condomínio" a categoria referência, e seja x_2 e x_3 as outras duas:

y = preço da casa

x_1 = metragem quadrada

x_2 = 1 se rancho, 0 caso contrário

x_3 = 1 se dois andares, 0 caso contrário

A equação de regressão múltipla é:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Interpretando os coeficientes de variáveis Dummy (com 3 níveis)

Considere a equação de regressão:

$$\hat{y} = 20.43 + 0.045x_1 + 23.53x_2 + 18.84x_3$$

Para um condomínio: $x_2 = x_3 = 0$

$$\hat{y} = 20.43 + 0.045x_1$$

Para um rancho: $x_2 = 1$; $x_3 = 0$

$$\hat{y} = 20.43 + 0.045x_1 + 23.53$$

Com os mesmos metros quadrados, uma fazenda terá preço médio estimado em 23,53 mil reais a mais que um condomínio

Para um andar duplo: $x_2 = 0$; $x_3 = 1$

$$\hat{y} = 20.43 + 0.045x_1 + 18.84$$

Com os mesmos metros quadrados, um sobrado terá um preço médio estimado de 18,84 mil reais a mais que um condomínio.

Desenho de Experimento

- Considere um experimento no qual quatro tratamentos serão usados, e o resultado também depende de três fatores ambientais que não podem ser controlados pelo experimentador
- Deixe a variável z_1 denotar o tratamento, onde $z_1 = 1, 2, 3, or\ 4$. Deixe z_2 denotar o fator de ambiente (a "variável de bloqueio"), onde $z_2 = 1, 2, or\ 3$

- Para modelar os quatro tratamentos, três variáveis dummies são necessárias
- Para modelar os três fatores ambientais, duas variáveis dummies são necessárias

Desenho de Experimento

(cont.)

- Defina cinco variáveis dummies, x_1, x_2, x_3, x_4, e x_5
- Deixe o nível de tratamento 1 ser o padrão ($z_1 = 1$)
  - Defina $x_1 = 1$ se $z_1 = 2$, $x_1 = 0$ caso contrário
  - Defina $x_2 = 1$ se $z_1 = 3$, $x_2 = 0$ caso contrário
  - Defina $x_3 = 1$ se $z_1 = 4$, $x_3 = 0$ caso contrário
- Deixe o nível de ambiente 1 ser o padrão ($z_2 = 1$)
  - Defina $x_4 = 1$ se $z_2 = 2$, $x_4 = 0$ caso contrário
  - Defina $x_5 = 1$ se $z_2 = 3$, $x_5 = 0$ caso contrário

Desenho de Experimento: Tabelas de variáveis Dummies

- Os valores das variáveis Dummies podem ser resumidos na tabela:

Z_1	X_1	X_2	X_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Z_2	X_4	X_5
1	0	0
2	1	0
3	0	1

Modelo de Desenho Experimental

- O Modelo de Desenho Experimental pode ser estimado usando a equação

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon$$

- O valor estimado de β_2 , por exemplo, mostra o total mostra a quantidade pela qual o valor y para o tratamento 3 excede o valor para o tratamento 1

Valores defasados da variável dependente

- Em modelos de series de tempo, os dados são coletados ao longo do tempo (semanal, trimestral, etc ...)
- O valor de y no período de tempo t é denotado y_t
- Os valores de y_t frequentemente depende do valor y_{t-1}, bem como outras variáveis independentes x_j :

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} \cdots + \beta_K x_{Kt} + \gamma y_{t-1} + \varepsilon_t$$

A variável dependente defasada é incluída como uma variável explicativa

Interpretando os Resultados dos Modelos Defasados

- Um aumento de 1 unidade na variável independente x_j no período de tempo t (todas as outras variáveis mantidas fixas), levará a um aumento esperado na variável dependente de
  - β_j no período t
  - $\beta_j \gamma$ no período (t+1)
  - $\beta_j \gamma^2$ no período (t+2)
  - $\beta_j \gamma^3$ no período (t+3) e assim por diante

- O aumento total esperado em todos os períodos de tempo atuais e futuros é $\beta_j/(1-\gamma)$

- Os coeficientes $\beta_0, \beta_1, \ldots, \beta_K, \gamma$

são estimados por mínimos quadrados da maneira usual

Interpretando os Resultados dos Modelos Defasados

(cont.)

- Intervalos de confiança e testes de hipótese para os coeficientes de regressão são calculados da mesma forma que na regressão múltipla ordinária

(Quando a equação de regressão contém variáveis defasadas, esses procedimentos são apenas aproximadamente válidos. A qualidade da aproximação melhora conforme o número de observações de amostra aumenta.)

Interpretando os Resultados dos Modelos Defasados

(cont.)

- Deve-se ter cuidado ao usar intervalos de confiança e testes de hipótese com dados de séries temporais

  - Existe a possibilidade de que os erros εi não sejam mais independentes um do outro.
  - Quando os erros são correlacionados, as estimativas dos coeficientes são imparciais, mas não eficientes. Assim, os intervalos de confiança e os testes de hipóteses não são mais válidos.

Viés de especificação

- Suponha que uma importante variável independente z seja omitida de um modelo de regressão
- Se z não estiver correlacionado com todas as outras variáveis independentes incluídas, a influência de z é deixada sem explicação e é absorvida pelo termo de erro, ε
- Mas se houver alguma correlação entre z e qualquer uma das variáveis independentes incluídas, parte da influência de z é capturada nos coeficientes das variáveis incluídas

Viés de especificação

(cont.)

- Se parte da influência da variável omitida z for capturada nos coeficientes das variáveis independentes incluídas, então esses coeficientes são enviesados ...
- ... e as declarações inferenciais usuais de teste de hipótese ou intervalos de confiança podem ser seriamente enganosas
- Além disso, o erro do modelo estimado incluirá o efeito das variáveis ausentes e será maior

Multicolinearidade

- Colinearidade: existe alta correlação entre duas ou mais variáveis independentes
- Isso significa que as variáveis correlacionadas contribuem com informações redundantes para o modelo de regressão múltipla

Multicolinearidade

(cont.)

Incluir duas variáveis explicativas altamente correlacionadas pode afetar adversamente os resultados da regressão

- Nenhuma nova informação fornecida
- Pode levar a coeficientes instáveis (grande erro padrão e baixos valores t)
- Sinais de coeficiente podem não corresponder às expectativas anteriores

Algumas indicações de forte multicolinearidade

- Sinais incorretos nos coeficientes
- Grande mudança no valor de um coeficiente anterior quando uma nova variável é adicionada ao modelo
- Uma variável anteriormente significativa se torna insignificante quando uma nova variável independente é adicionada
- A estimativa do desvio padrão do modelo aumenta quando uma variável é adicionada ao modelo

Detectando a Multicolinearidade

- Examine a matriz de correlação simples para determinar se existe forte correlação entre qualquer uma das variáveis independentes do modelo

- A multicolinearidade pode estar presente se o modelo parecer explicar bem a variável dependente (estatística F alta e baixa s_e) mas as estatísticas individuais do coeficiente t são insignificantes

Pressupostos da regressão

- Normalidade do Erro

Os valores de erro (ε) são normalmente distribuídos para qualquer valor de X

- Homocedasticidade

A distribuição de probabilidade dos erros tem variância constante

- Independência de Erros

Os valores de erro são estatisticamente independentes

Análise Residual

$$e_i = y_i - \hat{y}_i$$

- O residual para a observação i, e_i, é a diferença entre seu valor observado e predito pela regressão
- Verifique os pressupostos da regressão examinando os resíduos

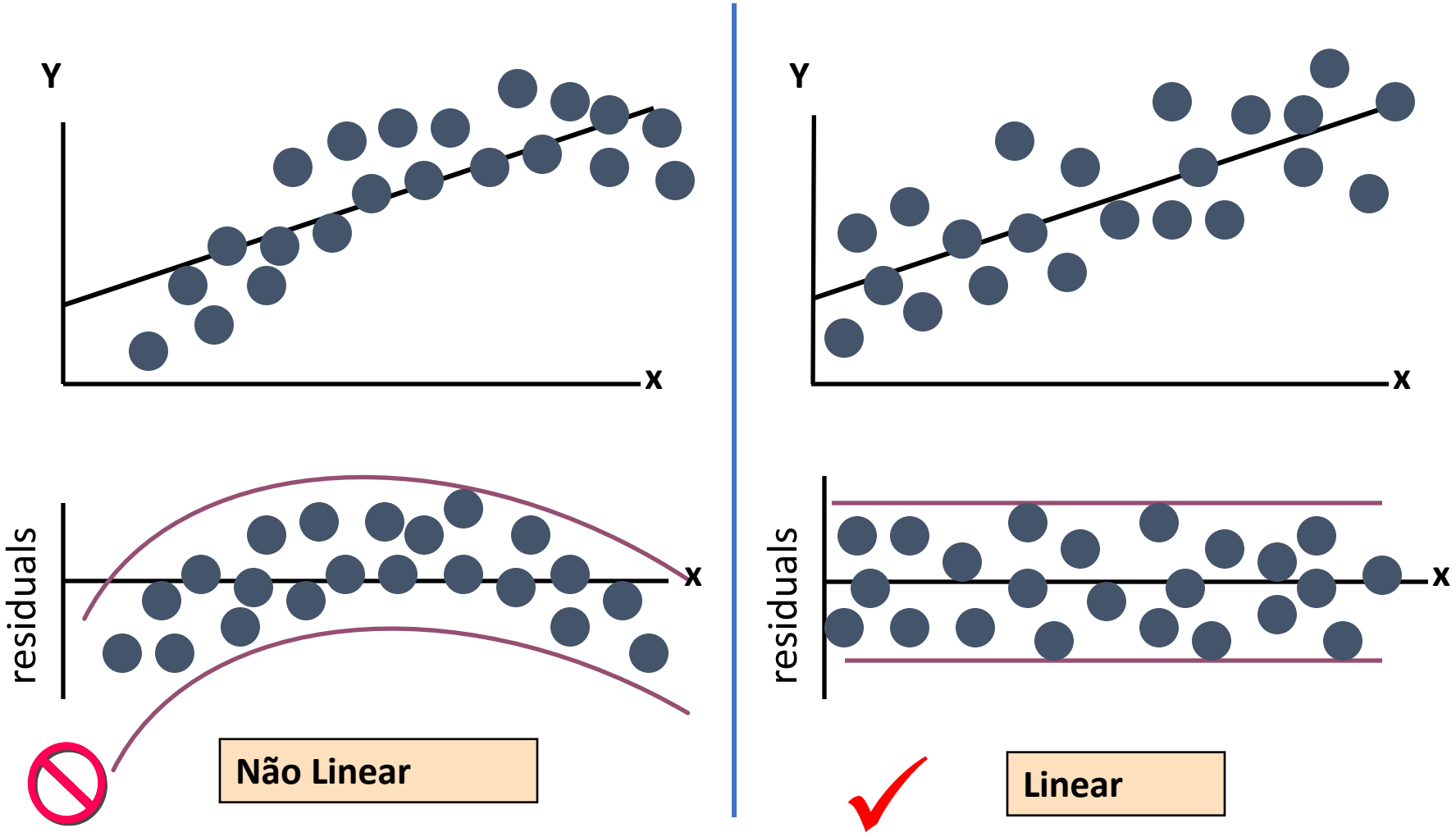
Examine o pressuposto de linearidade

- Examine a variação constante para todos os níveis de X (homocedasticidade)
- Avalie o pressuposto de distribuição normal
- Avalie o pressuposto de independência

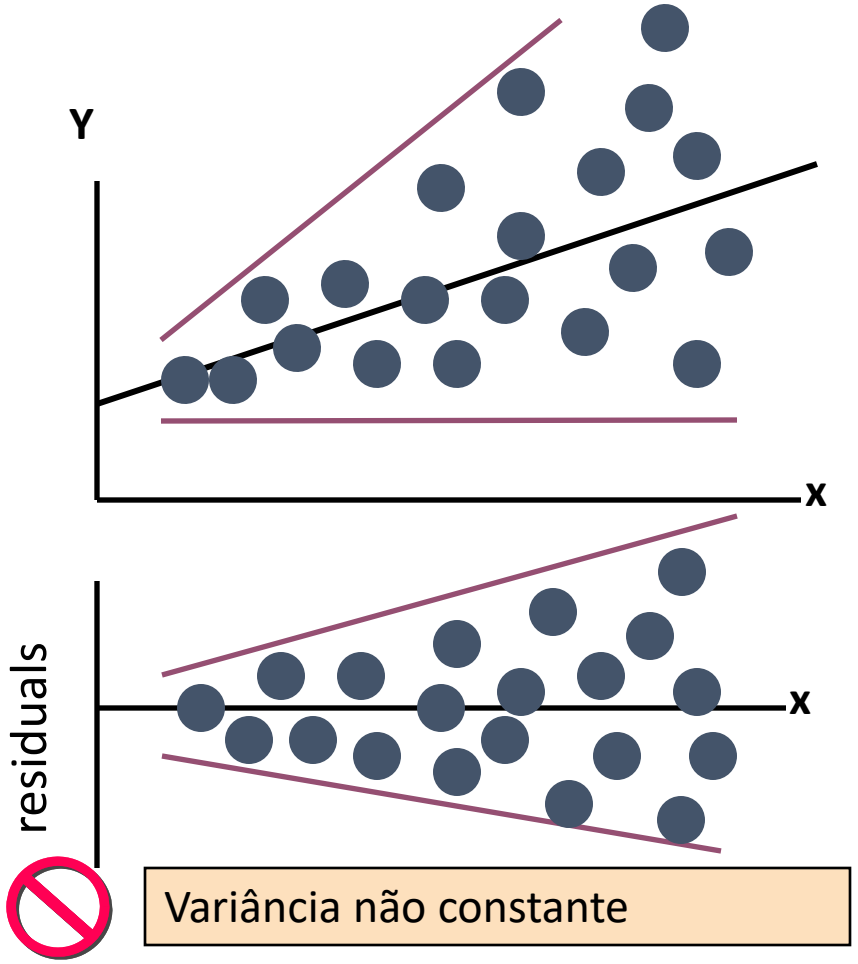
Análise gráfica de resíduos

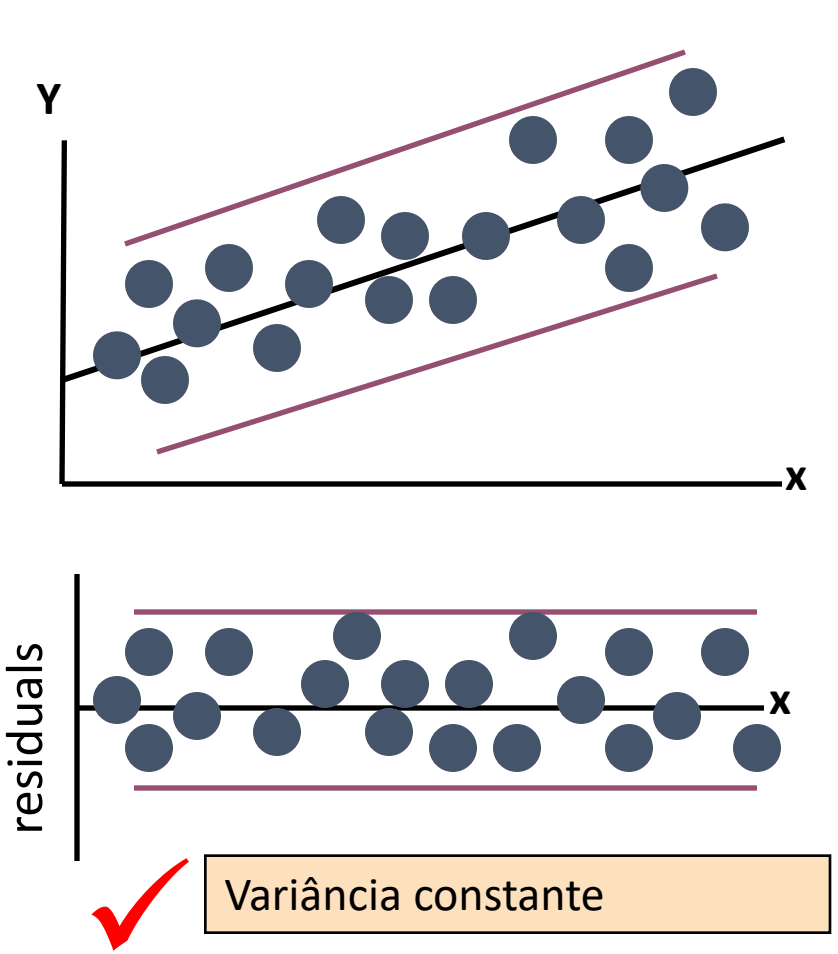
- Pode plotar resíduos vs. X

Análise Residual para Linearidade

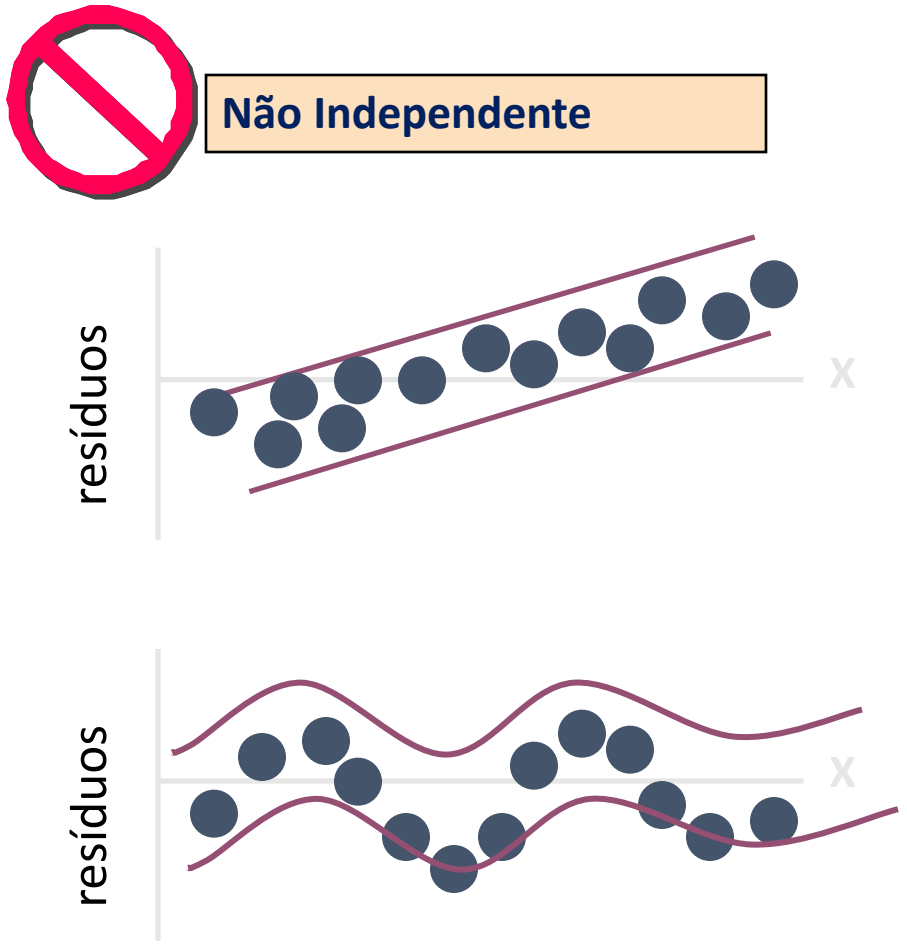


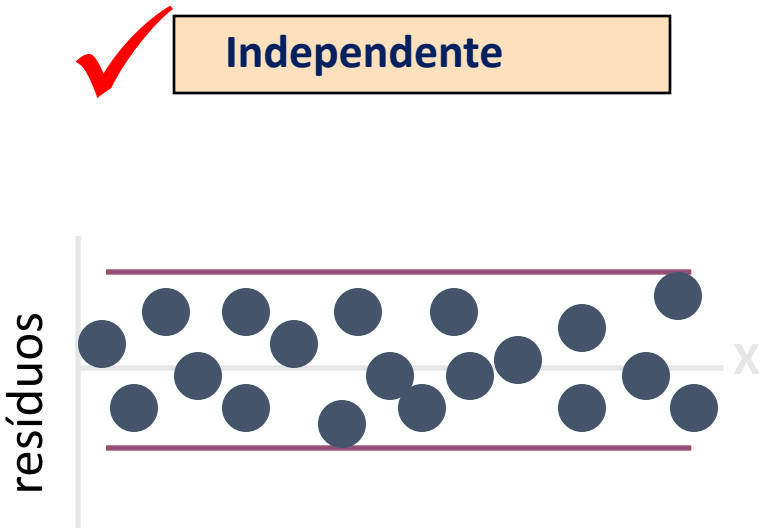
Análise Residual para Homoscedasticidade





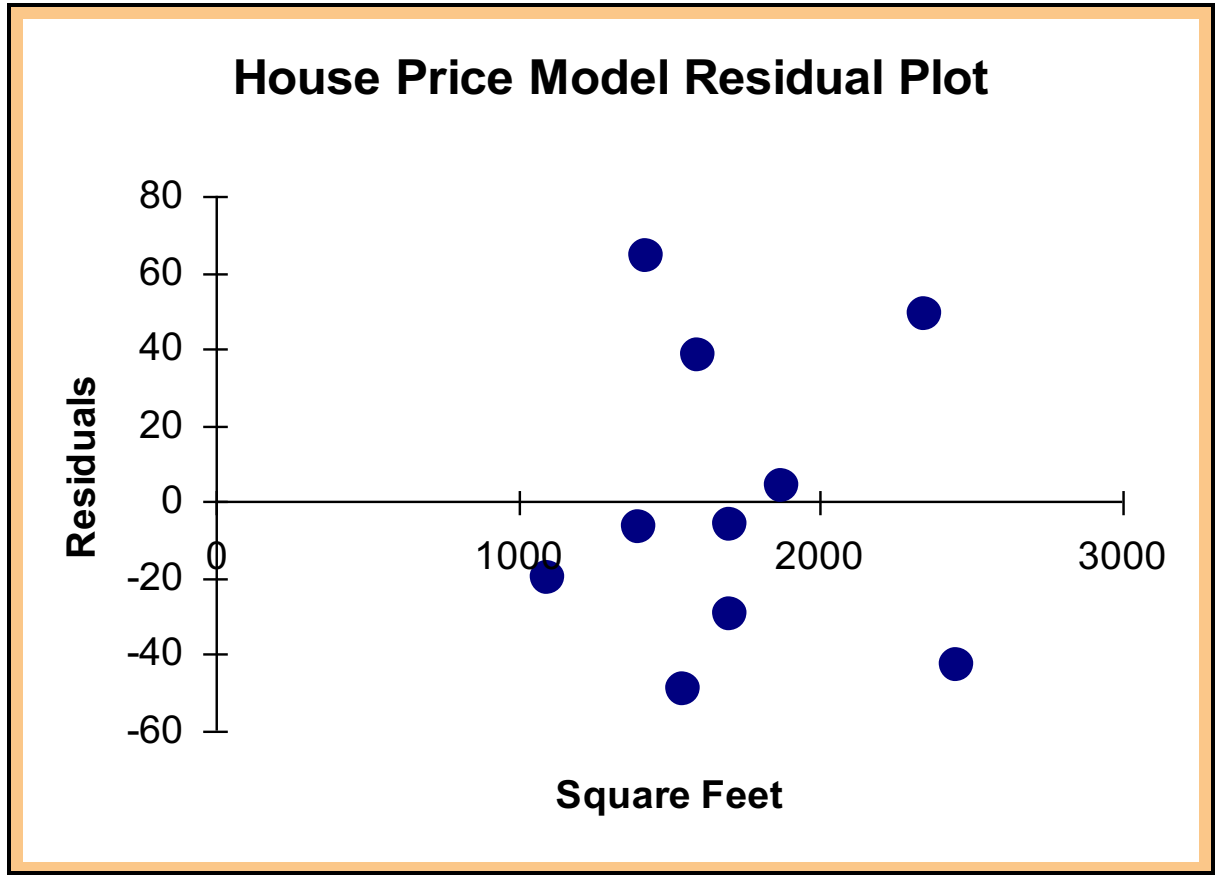
Análise Residual para Independência





Saída residual do Excel

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Não parece violar
quaisquer suposições de regressão

Heteroscedasticidade

- Homoscedasticidade
  - A distribuição de probabilidade dos erros tem variância constante
- Heteroscedasticidade
  - Os termos de erro não têm todos a mesma variação
  - O tamanho das variações de erro pode depender do tamanho do valor da variável dependente, por exemplo.

Heteroscedasticidade

(cont.)

- Quando a heteroscedasticidade está presente:
  - Mínimos quadrados não é o procedimento mais eficiente para estimar coeficientes de regressão
  - Os procedimentos usuais para derivar intervalos de confiança e testes de hipóteses não são válidos

Testes para a Heteroscedasticidade

- Para testar a hipótese nula de que os termos de erro, ε_i, todos têm a mesma variância contra a alternativa de que suas variâncias dependem dos valores esperados $\hat{y}_i$
- Estimar a regressão simples

$$e_i^2 = a_0 + a_1 \hat{y}_i$$

- Seja R^2 o coeficiente de determinação desta nova regressão

A hipótese nula é rejeitada se nR^2 é maior que $\chi^2_{1,\alpha}$

  - sendo $\chi^2_{1,\alpha}$ é o valor crítico da variável aleatória qui-quadrado com 1 grau de liberdade e probabilidade de erro α

Erros auto-correlacionados

Independência de Erros

- Os valores de erro são estatisticamente independentes

Erros autocorrelacionados

- Os resíduos em um período de tempo estão relacionados aos resíduos em outro período

Erros auto-correlacionados

(cont.)

A autocorrelação viola uma suposição de regressão de mínimos quadrados

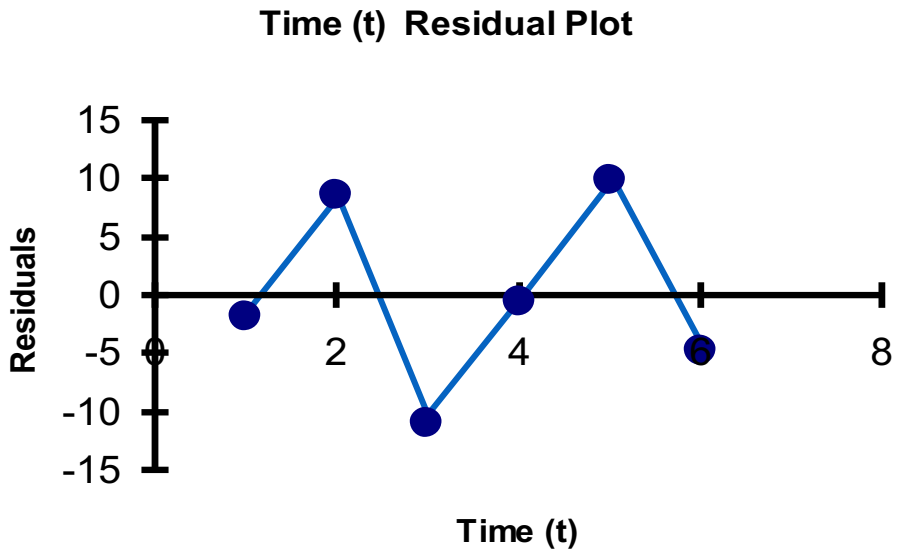
- Leva a estimativas muito pequenas de s_b (ou seja, tendenciosas)

- Assim, os valores de t são muito grandes e algumas variáveis podem parecer significativas quando não são

Autocorrelação

- A autocorrelação é a correlação dos erros (resíduos) ao longo do tempo

- Aqui, os resíduos mostram um padrão cíclico, não aleatório



- Viola o proessuposto de regressão de que os resíduos são aleatórios e independentes

A estatística Durbin-Watson

- A estatística Durbin-Watson é usada para testar a autocorrelação

H_0: resíduos sucessivos não são correlacionados
(i.e., $Corr(\varepsilon_t,\varepsilon_{t-1}) = 0$)

H_1: autocorrelação está presente

Como computer Durbin-Watson teste?

https://www.youtube.com/watch?v=J2DmtU4yu1o (excel)

https://www.youtube.com/watch?v=K3pCR1H3zls. (stata)

A estatística Durbin-Watson

H_0: $\rho = 0$ (não autocorrelação)

H_1: autocorrelação está presente

- A estatística de teste Durbin-Watson (d):

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- O intervalo possível é $0 \leq d \leq 4$
- d deve ser próximo a 2 se H_0 for verdadeiro
- d menor que 2 pode sinalizar autocorrelação positiva,
- d maior que 2 pode sinalizar autocorrelação negativa

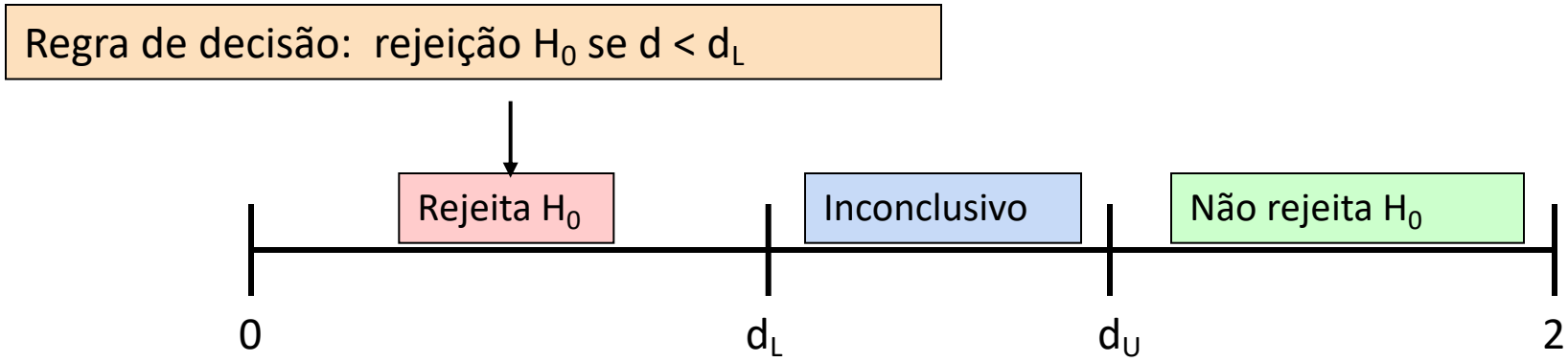
Teste de autocorrelação positiva

H_0: autocorrelação positiva não existe

H_1: autocorrelação positiva está presente

https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf

- Calcule a estatística de teste de Durbin-Watson = d
  - d pode ser aproximado por $d = 2(1 – r)$, sendo r é a correlação da amostra de erros sucessivos
- Encontre os valores d_L e d_U na tabela de Durbin-Watson
  - (para amostra de tamanho n e número de variáveis independentes K)



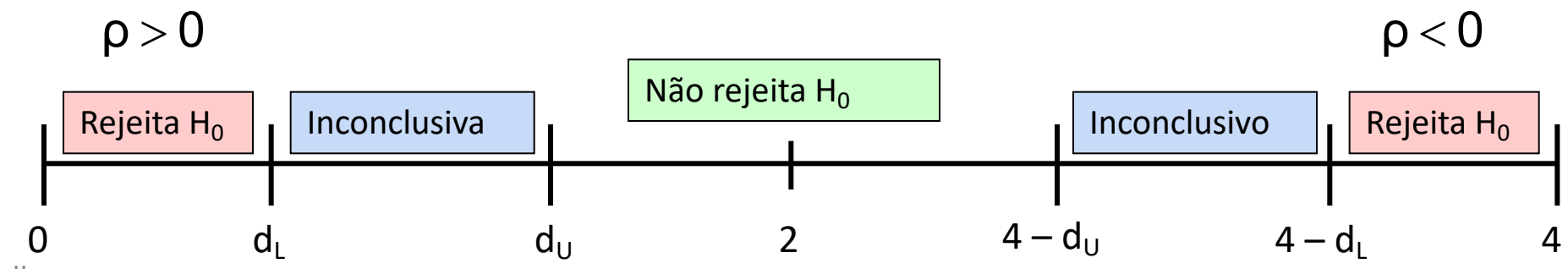
Autocorrelação Negativa

Existe autocorrelação negativa se erros sucessivos forem correlacionados negativamente

Isso pode ocorrer se erros sucessivos se alternarem no sinal

Regra de Decisão para autocorrelação negativa:

rejeita H_0 se $d > 4 - d_L$

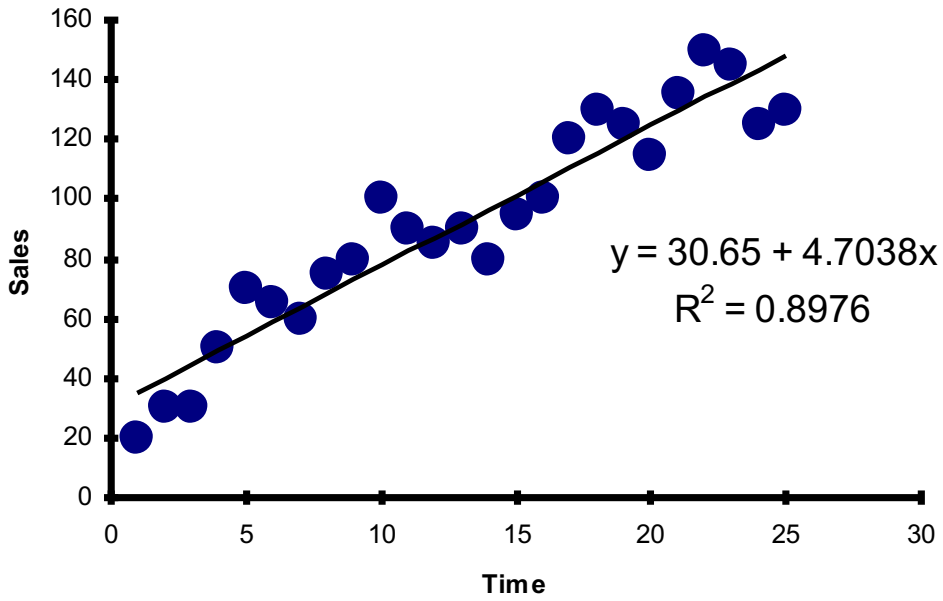


Testando a Autocorrelação positiva

(cont.)

- Exemplo com n = 25:

Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**



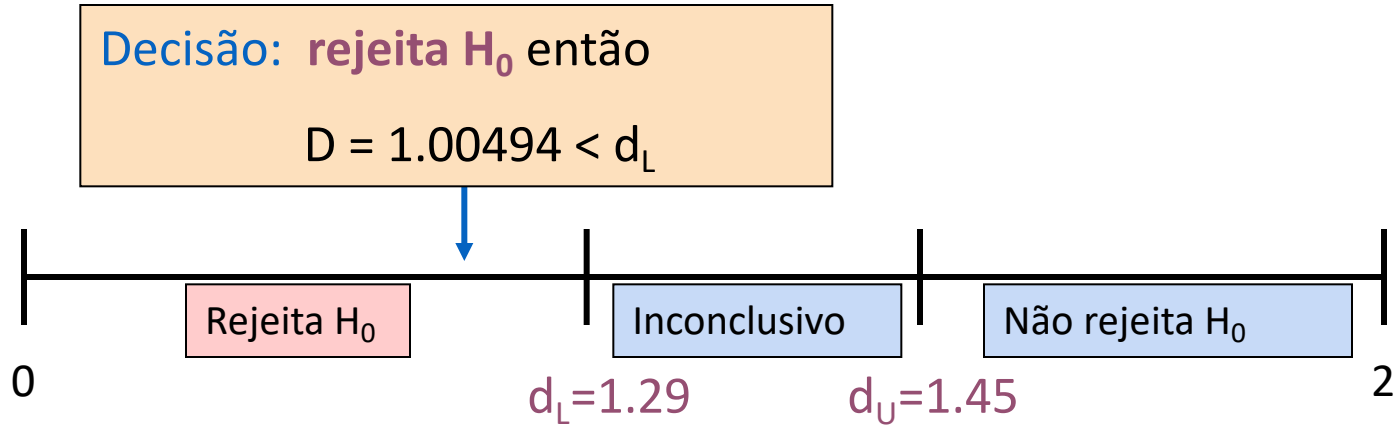
$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{3296.18}{3279.98} = 1.00494$$

Testando a Autocorrelação positiva

(cont.)

- Aqui, n = 25 e existe k = 1 variáveis independentes
- Usando a tabela Durbin-Watson, $d_L = 1.29$ e $d_U = 1.45$
- $D = 1.00494 < d_L = 1.29$, então rejeita H_0 e conclua que a autocorrelação positiva significante existe.

Portanto, o modelo linear não é o modelo apropriado para prever vendas



Lidando com Autocorrelação

- Suponha que queremos estimar os coeficientes do modelo de regressão

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

sendo que o termo erro ε_t é autocorrelacionado

- Dois passos:

(i) Estime o modelo por mínimos quadrados, obtendo a estatística de Durbin-Watson, d, e então estime o parâmetro de autocorrelação usando

$$r = 1 - \frac{d}{2}$$

Lidando com Autocorrelação

(ii) Estimar por mínimos quadrados uma segunda regressão com

- Variável dependente $(y_t - ry_{t-1})$
- Variáveis independentes $(x_{1t} - rx_{1,t-1})$, $(x_{2t} - rx_{2,t-1})$, . . ., $(x_{k1t} - rx_{k,t-1})$

- Os parâmetros $\beta_1, \beta_2, \ldots, \beta_k$ são coeficientes de regressão estimados do segundo modelo
- Uma estimativa de β_0 é obtido dividindo a interceptação estimada para o segundo modelo por(1-r)
- Testes de hipótese e intervalos de confiança para os coeficientes de regressão podem ser realizados usando a saída do segundo modelo