ANÁLISE DE DADOS EM GRH

Aula 7:

'Que fatores explicam a variação nos salários na organização?

Validação de Modelos de Regressão Linear

Docente: Daniela Craveiro
dcraveiro@iseg.ulisboa.pt

ANÁLISE DE DADOS EM GRH

- **Na Aula Anterior**
  - **Aprendemos a implementar e interpretar o resultado de um modelo de regressão linear**
- **Objetivos da Aula**
  - **Parte Teórica**
    - **Perceber qual a necessidade de fazermos diagnósticos aos pressupostos do nosso modelo de regressão**
    - **Saber quais são os pressupostos do modelo de regressão linear**
    - **Saber como, com a ajuda de gráficos e testes estatísticos, podemos conferir se os pressupostos do modelo estão a ser cumpridos**
  - **Parte Prática**
    - **Saber implementar os estudo dos pressupostos do modelo no SPSS**

Pressupostos do modelo de regressão

	PRESSUPOSTOS	DEFINIÇÃO	FORMA DE VALIDAÇÃO
I	Linearidade	O efeito das variáveis independentes na variável dependente é linear e aditivo.	• Análise gráfica (Matriz de Dispersão, por exemplo)

Pressupostos do modelo de regressão

	PRESSUPOSTOS	DEFINIÇÃO	FORMA DE VALIDAÇÃO
I	Linearidade	O efeito das variáveis independentes na variável dependente é linear e aditivo.	• Análise gráfica (Matriz de Dispersão, por exemplo)
II	Normalidade da Distribuição dos Erros	Os erros seguem uma distribuição normal.	• Análise de Resíduos • Gráfico de Q-Q

Pressupostos do modelo de regressão

	PRESSUPOSTOS	DEFINIÇÃO	FORMA DE VALIDAÇÃO
I	Linearidade	O efeito das variáveis independentes na variável dependente é linear e aditivo.	• Análise gráfica (Matriz de Dispersão, por exemplo)
II	Normalidade da Distribuição dos Erros	Os erros seguem uma distribuição normal.	• Análise de Resíduos • Gráfico de Q-Q
III	Média Condicional Zero dos Erros	O termo de erro aleatório tem valor esperado igual a zero.	• Análise de Resíduos

Pressupostos do modelo de regressão

	PRESSUPOSTOS	DEFINIÇÃO	FORMA DE VALIDAÇÃO
I	Linearidade	O efeito das variáveis independentes na variável dependente é linear e aditivo.	• Análise gráfica (Matriz de Dispersão, por exemplo)
II	Normalidade da Distribuição dos Erros	Os erros seguem uma distribuição normal.	• Análise de Resíduos • Gráfico de Q-Q
III	Média Condicional Zero dos Erros	O termo de erro aleatório tem valor esperado igual a zero.	• Análise de Resíduos
IV	Homocedasticidade (ou Igual Variância)	A distribuição dos erros apresenta uma variância constante (hipótese da homocedasticidade).	• Análise de Resíduos

Pressupostos do modelo de regressão

	PRESSUPOSTOS	DEFINIÇÃO	FORMA DE VALIDAÇÃO
I	Linearidade	O efeito das variáveis independentes na variável dependente é linear e aditivo.	• Análise gráfica (Matriz de Dispersão, por exemplo)
II	Normalidade da Distribuição dos Erros	Os erros seguem uma distribuição normal.	• Análise de Resíduos • Gráfico de Q-Q
III	Média Condicional Zero dos Erros	O termo de erro aleatório tem valor esperado igual a zero.	• Análise de Resíduos
IV	Homocedasticidade (ou Igual Variância)	A distribuição dos erros apresenta uma variância constante (hipótese da homocedasticidade).	• Análise de Resíduos
V	Independência dos Erros	Os erros não estão correlacionados, i.e., o valor de um erro não depende de qualquer outro erro.	• Dublin-Watson

Pressupostos do modelo de regressão

	PRESSUPOSTOS	DEFINIÇÃO	FORMA DE VALIDAÇÃO
I	Linearidade	O efeito das variáveis independentes na variável dependente é linear e aditivo.	• Análise gráfica (Matriz de Dispersão, por exemplo)
II	Normalidade da Distribuição dos Erros	Os erros seguem uma distribuição normal.	• Análise de Resíduos • Gráfico de Q-Q
III	Média Condicional Zero dos Erros	O termo de erro aleatório tem valor esperado igual a zero.	• Análise de Resíduos
IV	Homocedasticidade (ou Igual Variância)	A distribuição dos erros apresenta uma variância constante (hipótese da homocedasticidade).	• Análise de Resíduos
V	Independência dos Erros	Os erros não estão correlacionados, i.e., o valor de um erro não depende de qualquer outro erro.	• Dublin-Watson
VI	Ausência de multicolinearidade perfeita	As variáveis independents não estão perfeitamente correlacionadas entre si.	• Diagnósticos de Colinearidade

Pressupostos do modelo de regressão

	PRESSUPOSTOS	DEFINIÇÃO	FORMA DE VALIDAÇÃO
I	Linearidade	O efeito das variáveis independentes na variável dependente é linear e aditivo.	• Análise gráfica (Matriz de Dispersão, por exemplo)
II	Normalidade da Distribuição dos Erros	Os erros seguem uma distribuição normal.	• Análise de Resíduos • Gráfico de Q-Q
III	Média Condicional Zero dos Erros	O termo de erro aleatório tem valor esperado igual a zero.	• Análise de Resíduos
IV	Homocedasticidade (ou Igual Variância)	A distribuição dos erros apresenta uma variância constante (hipótese da homocedasticidade).	• Análise de Resíduos
V	Independência dos Erros	Os erros não estão correlacionados, i.e., o valor de um erro não depende de qualquer outro erro.	• Dublin-Watson
VI	Ausência de multicolinearidade perfeita	As variáveis independents não estão perfeitamente correlacionadas entre si.	• Diagnósticos de Colinearidade
VII	Ausência de Observações Influentes	Não existem observações que tenham uma influência anormal nos resultados do modelo.	• Cook's Distance

E qual é o problema se estes pressupostos não se verificarem?

- **Os intervalos de confiança ou os p-values podem estar a ser subestimados (i.e. mais pequenos do que na realidade são) …**

ou seja: estamos a atribuir significância estatística a uma estimativa que na realidade não a terá!

Como podemos saber se estes pressupostos estão a ser cumpridos?

Validação do Modelo de Regressão Linear

1. Estimar o modelo de regressão com os diagnósticos

2. Avaliação do Pressuposto II: Normalidade da Distribuição dos Erros

3. Avaliação do Pressuposto III: Média Condicional Zero dos Erros

4. Avaliação do Pressuposto IV: Homocedasticidade

4. Avaliação do Pressuposto V: Independência dos Erros

5. Avaliação do Pressuposto VI: Ausência de Multicolinearidade Perfeita

6. Avaliação do Pressuposto VII: Ausência de Observações Influentes

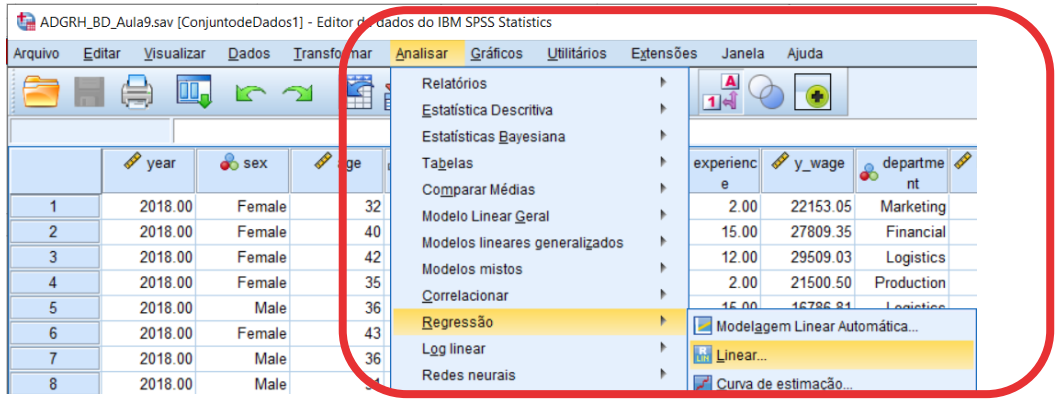
Validação do Modelo de Regressão Linear

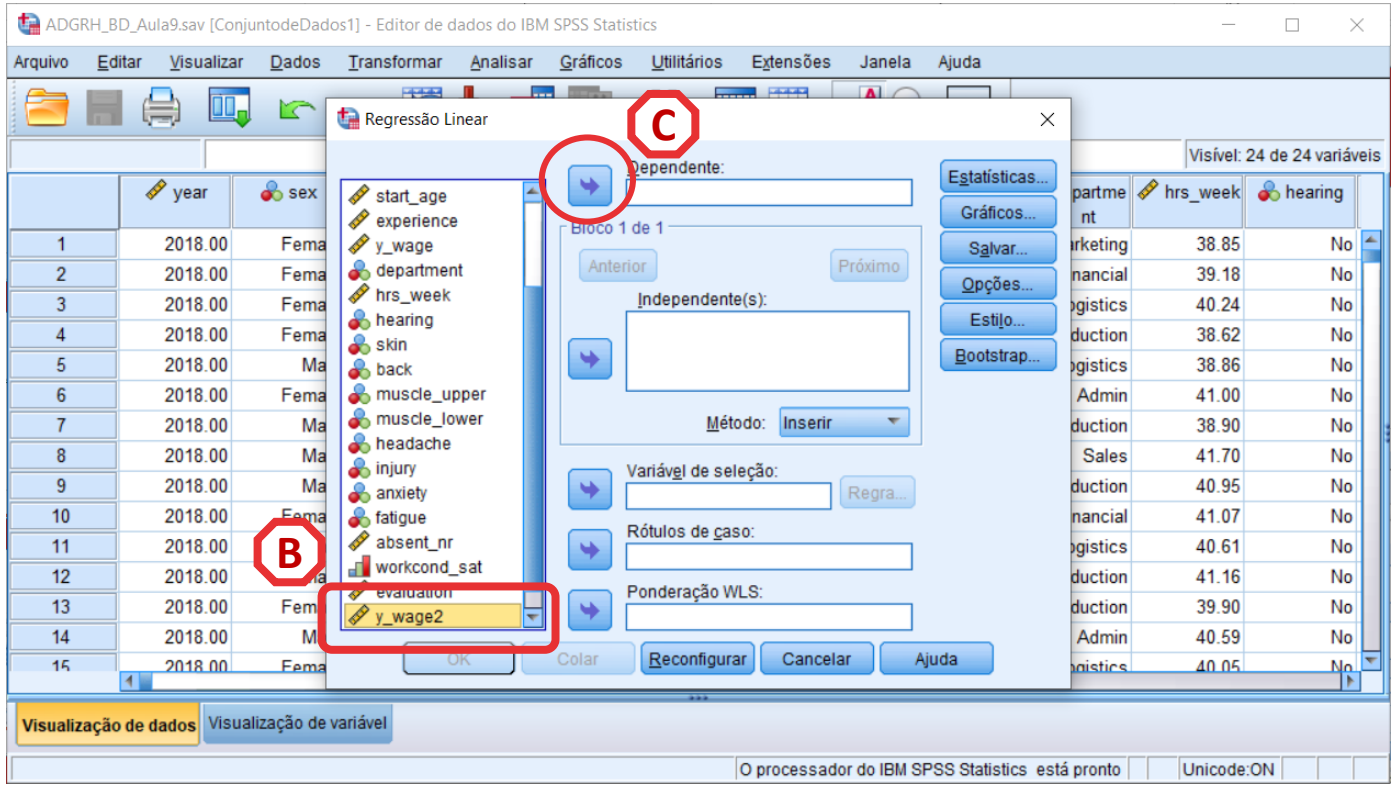
1. Estimar o modelo de regressão com os diagnósticos

Diagnósticos

- **Selecionar ‘Analisar’ / ‘Regressão’ / ‘Linear’**
- **Selecionar a variável ‘y_wage2’**
- **Colocar na caixa ‘Dependente’**

<u>Exercício</u>: Colocar as variáveis ‘sex_female’, ‘experience’ e ‘evaluation’ na caixa ‘Independente(s)’



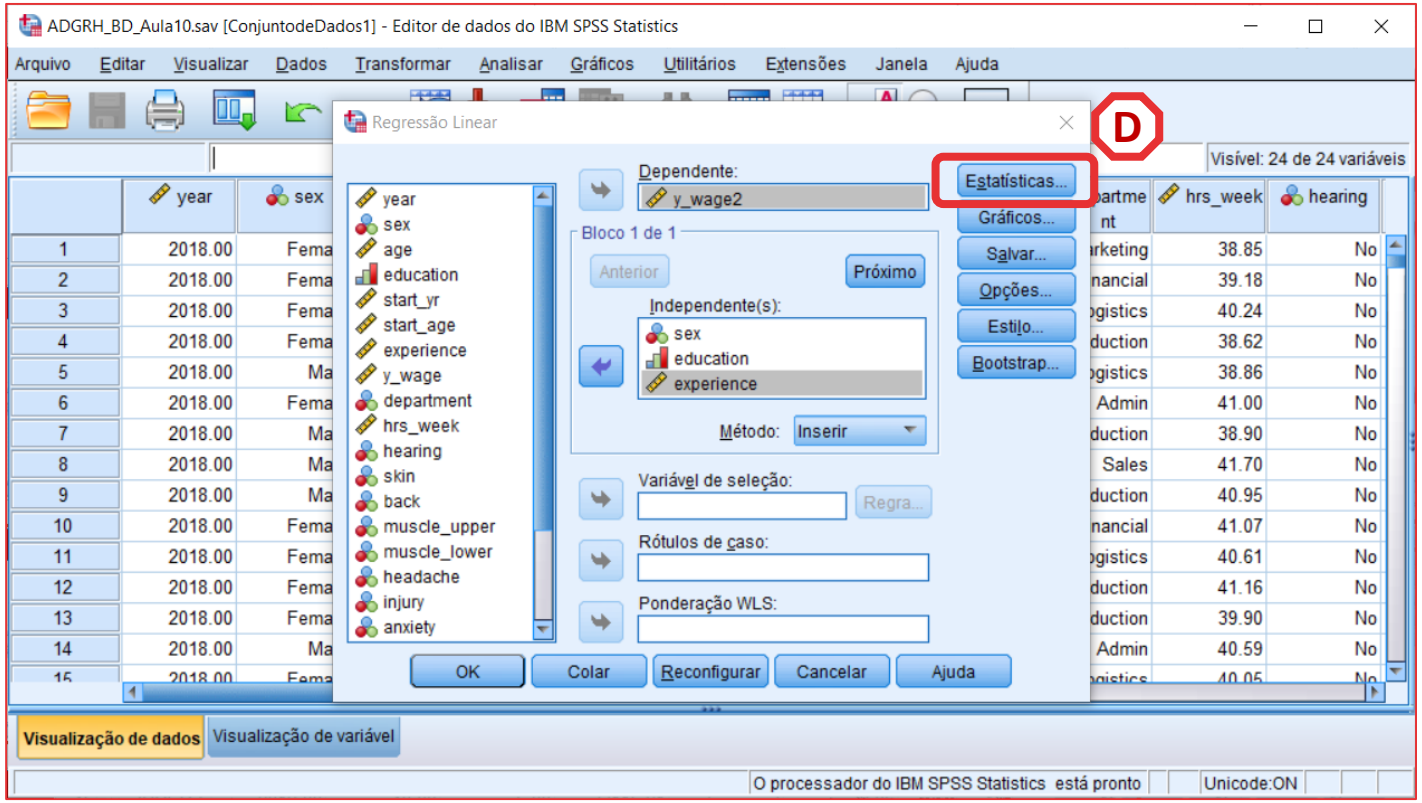


Diagnósticos

- Selecionar ‘Analisar’ / ‘Regressão’ / ‘Linear’ (A)
- Selecionar a variável ‘y_wage2’ (B)
- Colocar na caixa ‘Dependente’ (C)

  <u>Exercício</u>: Colocar as variáveis ‘sex_female’, ‘experience’ e ‘evaluation’ na caixa ‘Independente(s)’

- Selecionar botão ‘Estatísticas’ (D)

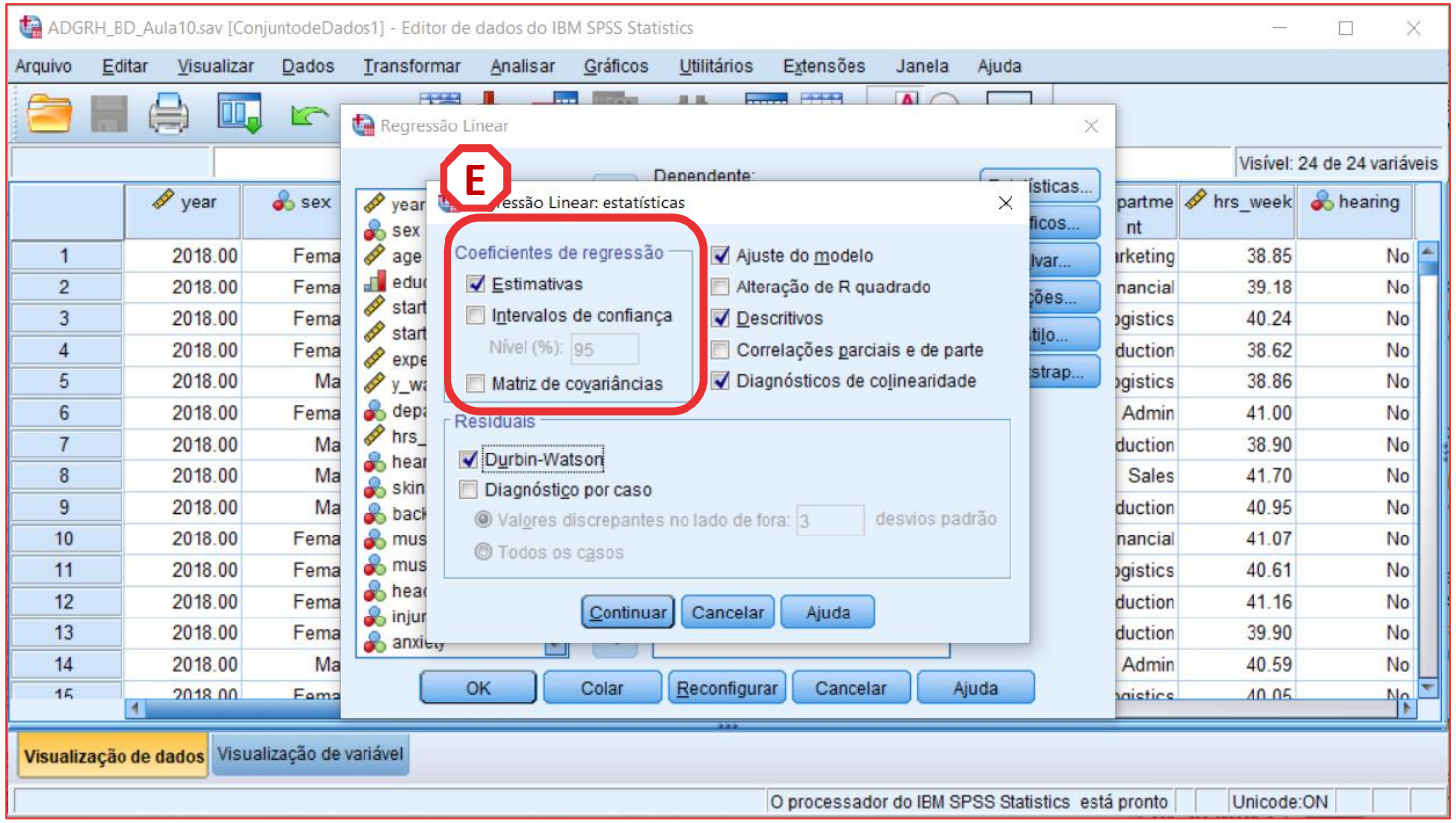


Diagnósticos

- Selecionar ‘Analisar’ / ‘Regressão’ / ‘Linear’ (A)
- Selecionar a variável ‘y_wage2’ (B)
- Colocar na caixa ‘Dependente’ (C)

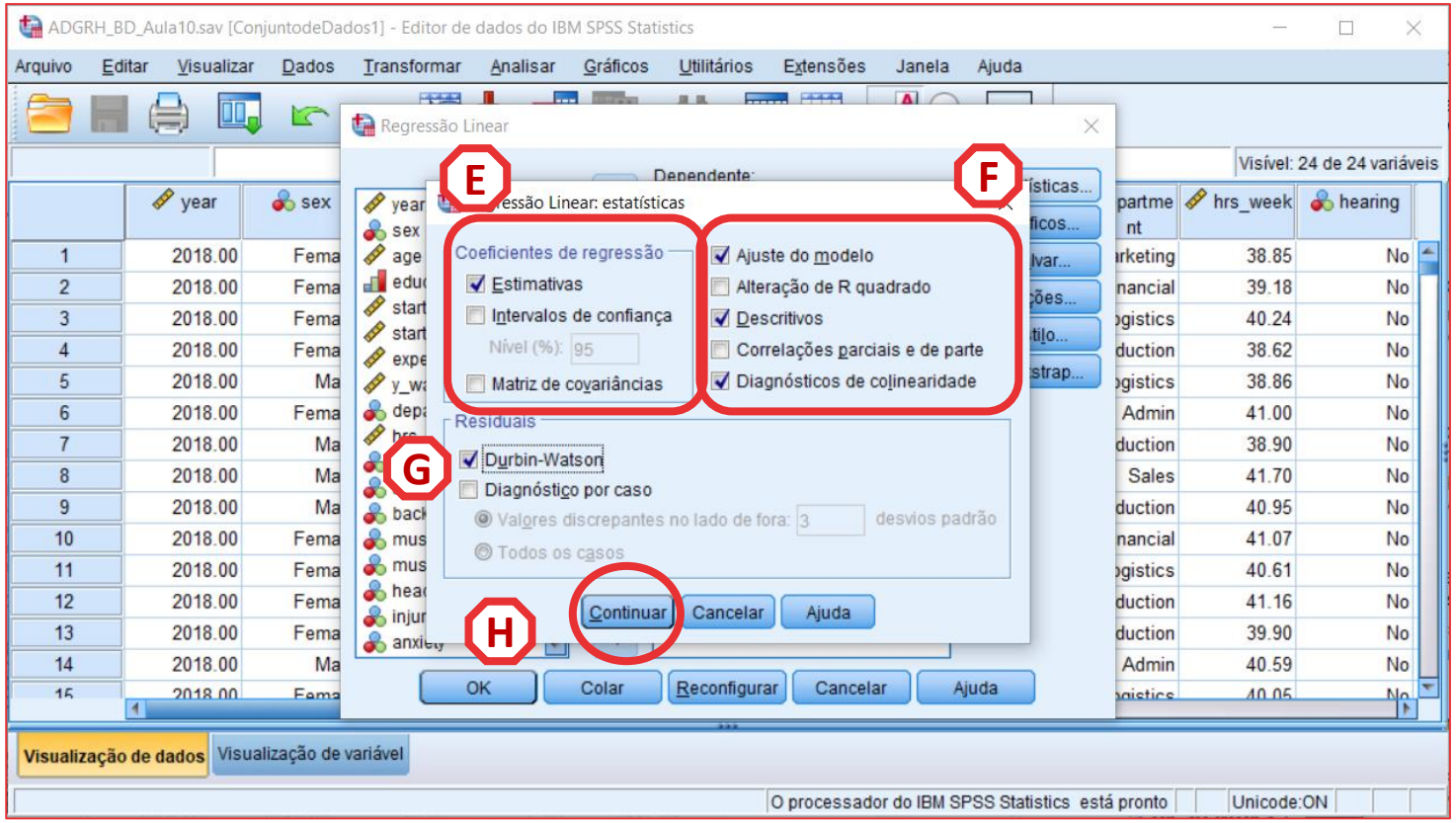
  Exercício: Colocar as variáveis ‘sex_female’, ‘experience’ e ‘evaluation’ na caixa ‘Independente(s)’

- Selecionar botão ‘Estatísticas’ (D)
- Selecionar ‘Estimativas’ (E)



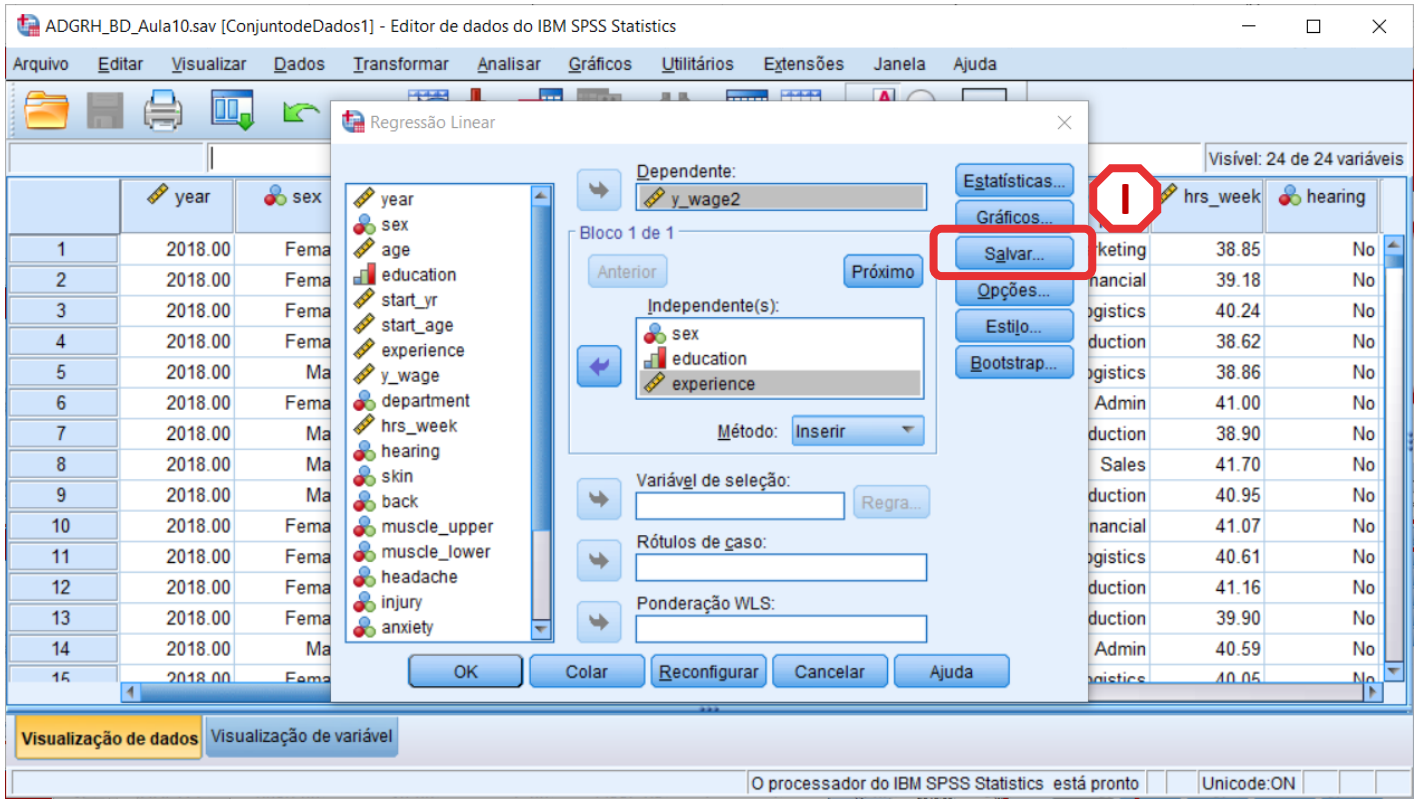
Diagnósticos

- Selecionar 'Ajuste do modelo' (F)
- Selecionar 'Descritivos'
- Selecionar 'Diagósticos de colinearidade'
- Selecionar 'Dublin-Watson' (G)
- Selecionar 'Continuar' (H)



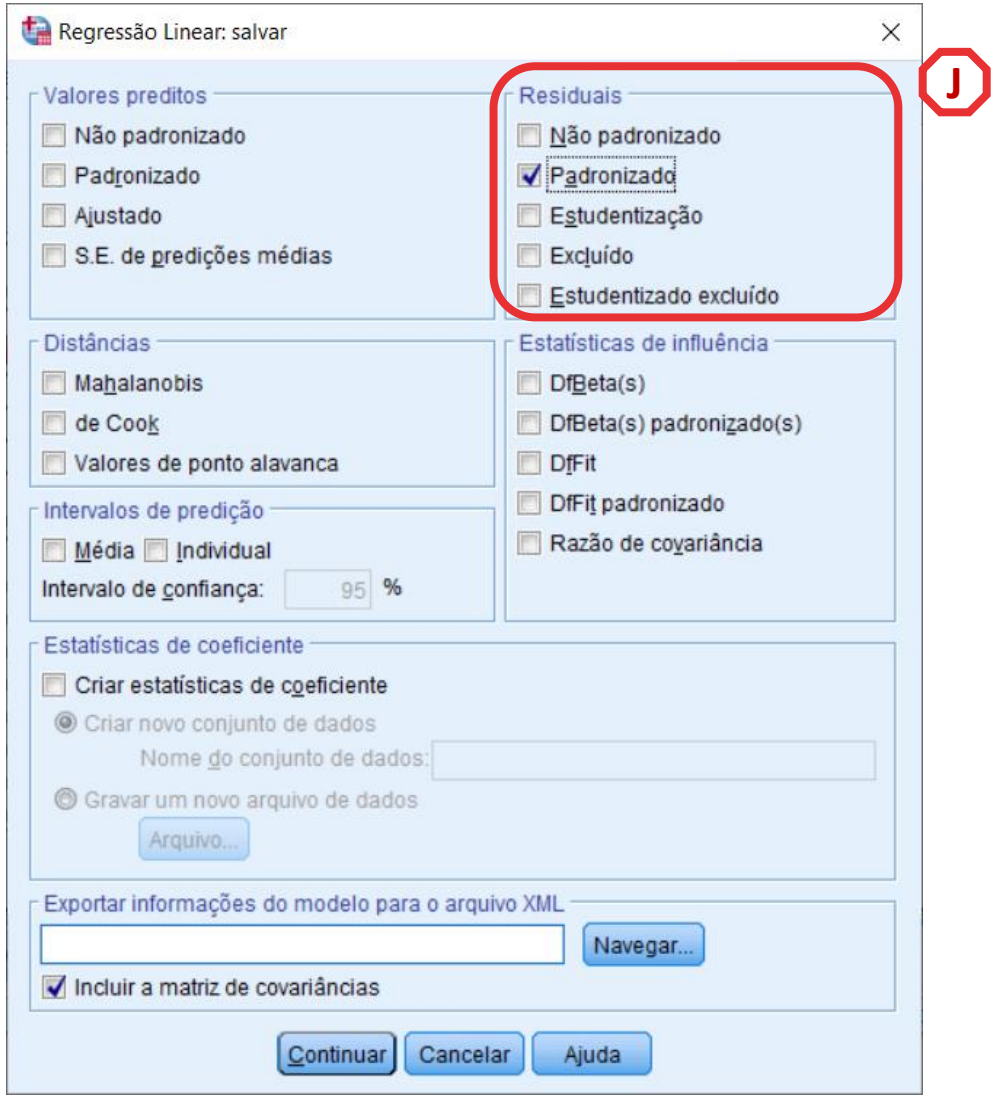
Diagnósticos

- **Selecionar botão ‘Salvar’**



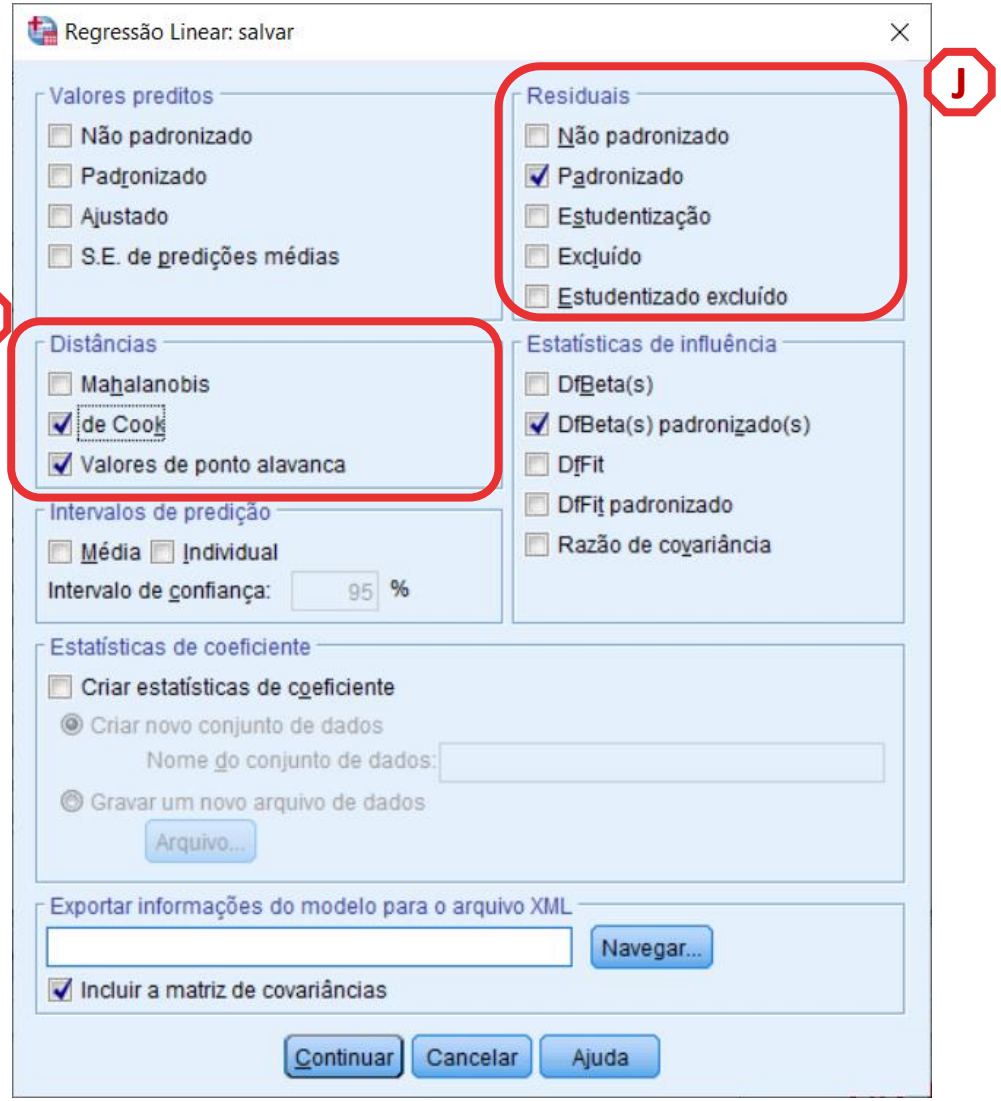
Diagnósticos

- **Selecionar botão ‘Salvar’** I
- **Selecionar ‘Padronizado’** J



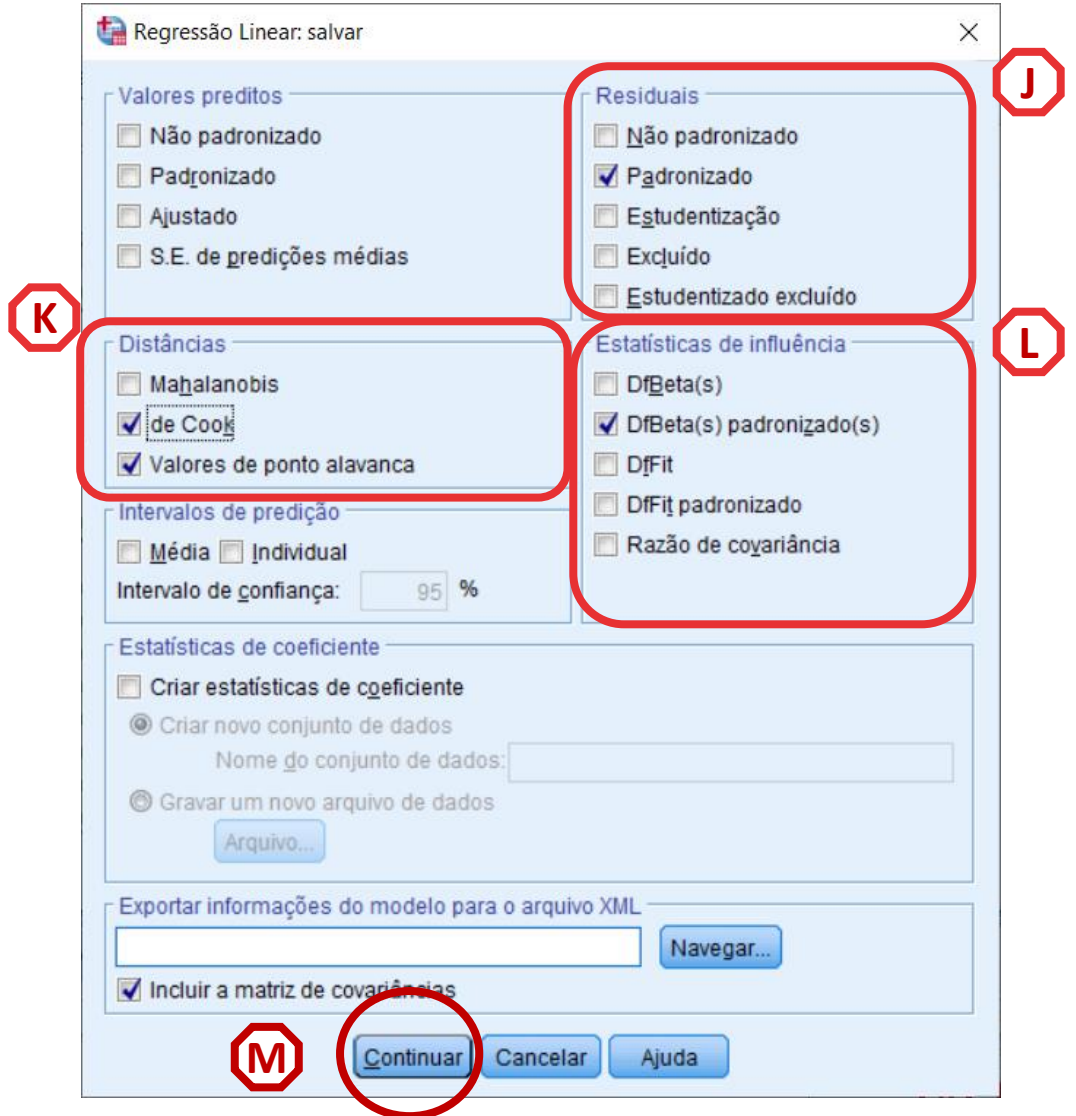
Diagnósticos

- **Selecionar botão ‘Salvar’** (I)
- **Selecionar ‘Padronizado’** (J)
- **Selecionar ‘de Cook’ e ‘Valores de ponto alavanca’** (K)



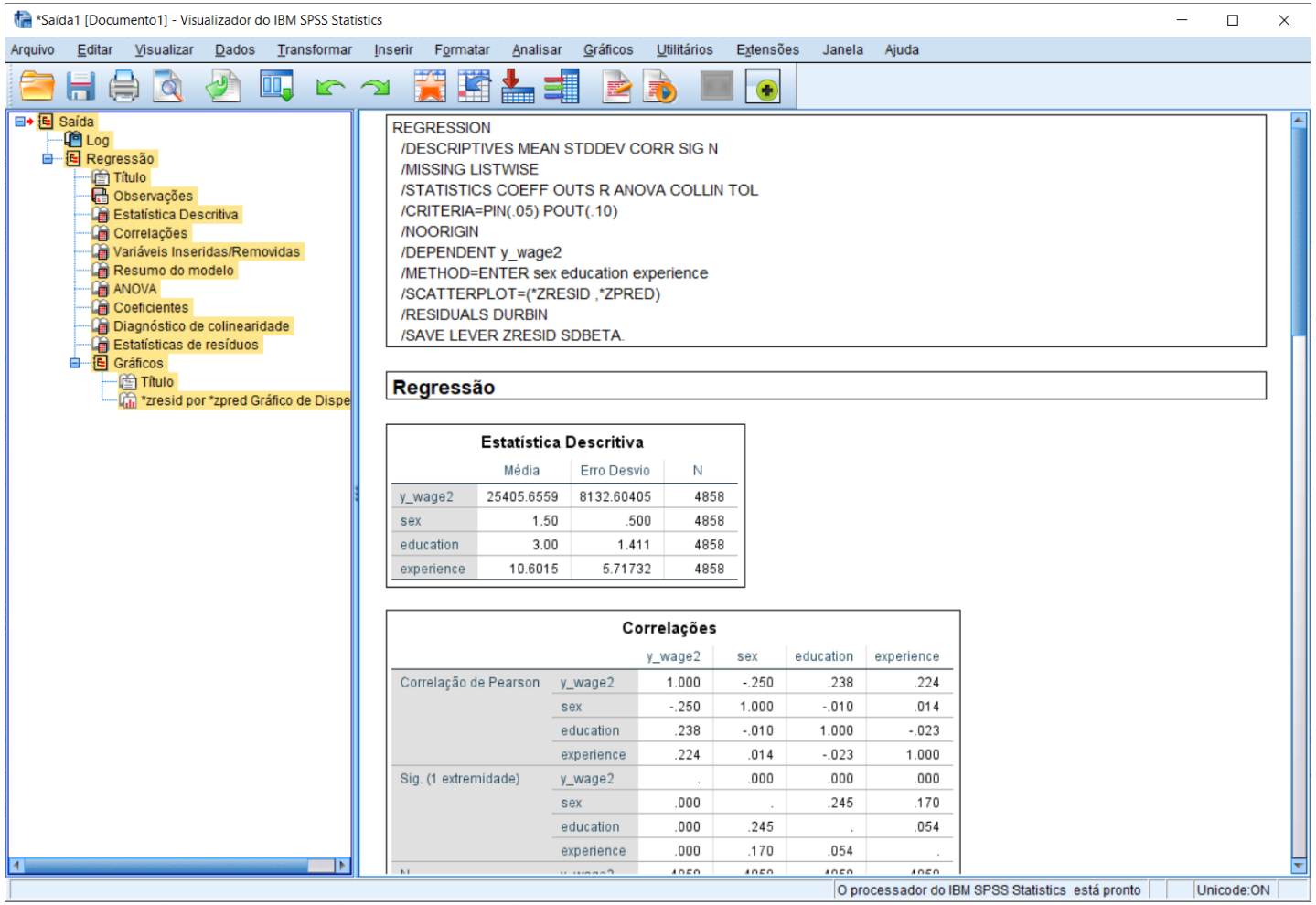
Diagnósticos

- Selecionar botão ‘Salvar’ (I)
- Selecionar ‘Padronizado’ (J)
- Selecionar ‘de Cook’ e ‘Valores de ponto alavanca’ (K)
- Selecionar ‘DfBeta(s) padronizado(s) (L)
- Selecionar ‘Continuar’/ ‘OK’ (M)



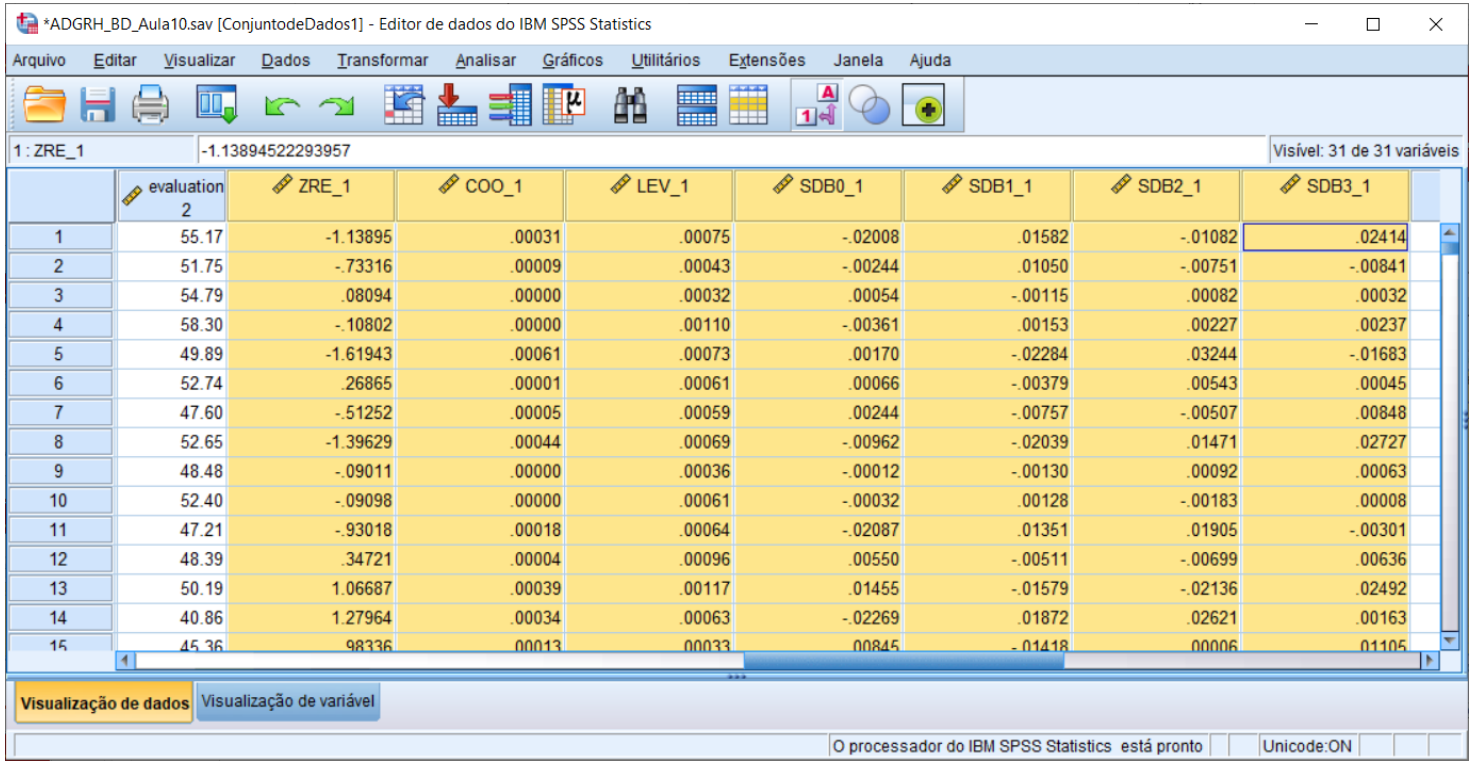
Diagnósticos

- **Os resultados são publicados no ‘Visualizador de Resultados’**



Diagnósticos

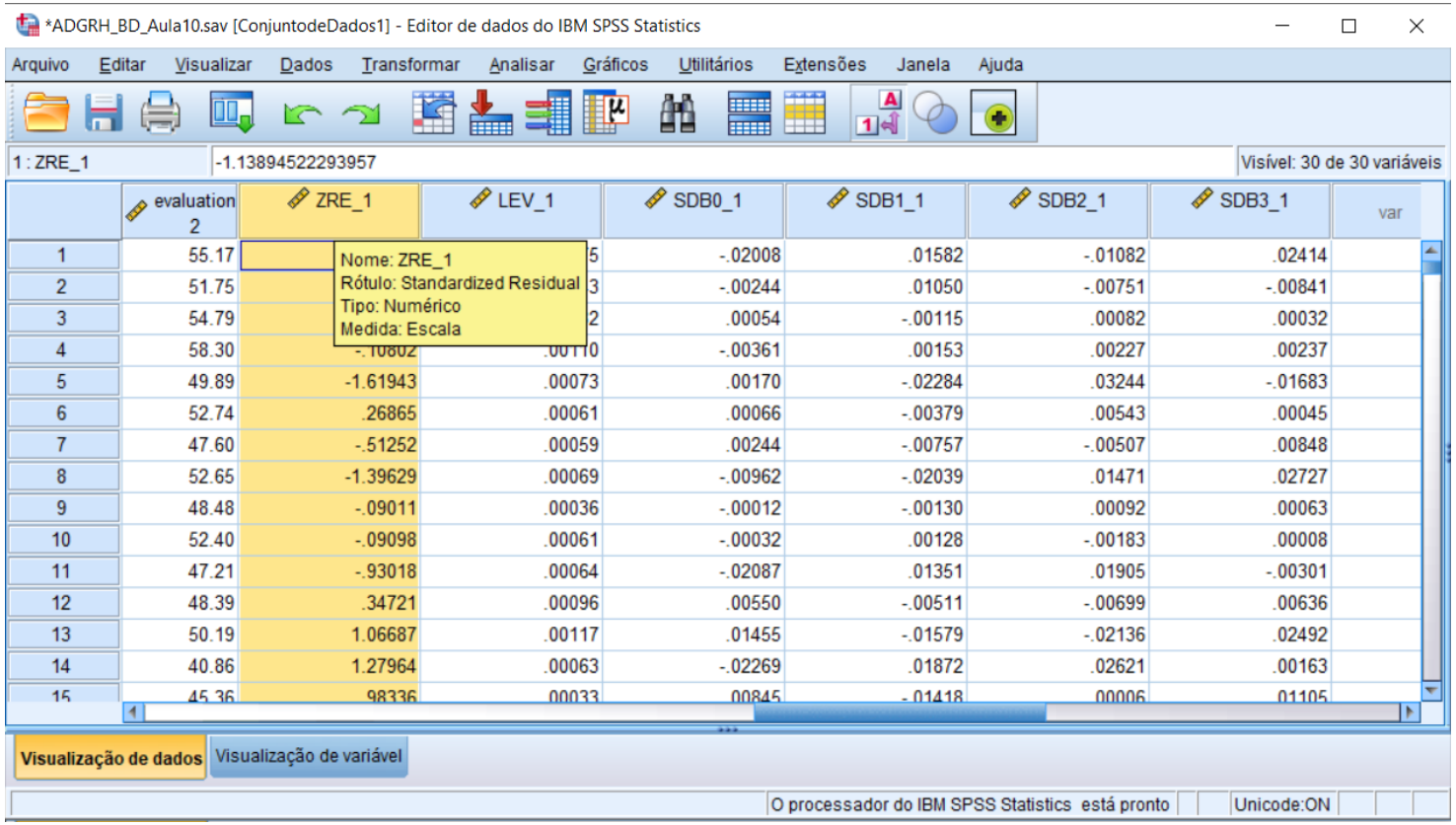
- **Quando instruímos o SPSS para produzir os diagnósticos, é criado um conjunto variáveis**



	evaluation 2	ZRE_1	COO_1	LEV_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1
1	55.17	-1.13895	.00031	.00075	-.02008	.01582	-.01082	.02414
2	51.75	-.73316	.00009	.00043	-.00244	.01050	-.00751	-.00841
3	54.79	.08094	.00000	.00032	.00054	-.00115	.00082	.00032
4	58.30	-.10802	.00000	.00110	-.00361	.00153	.00227	.00237
5	49.89	-1.61943	.00061	.00073	.00170	-.02284	.03244	-.01683
6	52.74	.26865	.00001	.00061	.00066	-.00379	.00543	.00045
7	47.60	-.51252	.00005	.00059	.00244	-.00757	-.00507	.00848
8	52.65	-1.39629	.00044	.00069	-.00962	-.02039	.01471	.02727
9	48.48	-.09011	.00000	.00036	-.00012	-.00130	.00092	.00063
10	52.40	-.09098	.00000	.00061	-.00032	.00128	-.00183	.00008
11	47.21	-.93018	.00018	.00064	-.02087	.01351	.01905	-.00301
12	48.39	.34721	.00004	.00096	.00550	-.00511	-.00699	.00636
13	50.19	1.06687	.00039	.00117	.01455	-.01579	-.02136	.02492
14	40.86	1.27964	.00034	.00063	-.02269	.01872	.02621	.00163
15	45.36	.98336	.00013	.00033	.00845	-.01418	.00006	.01105

Diagnósticos

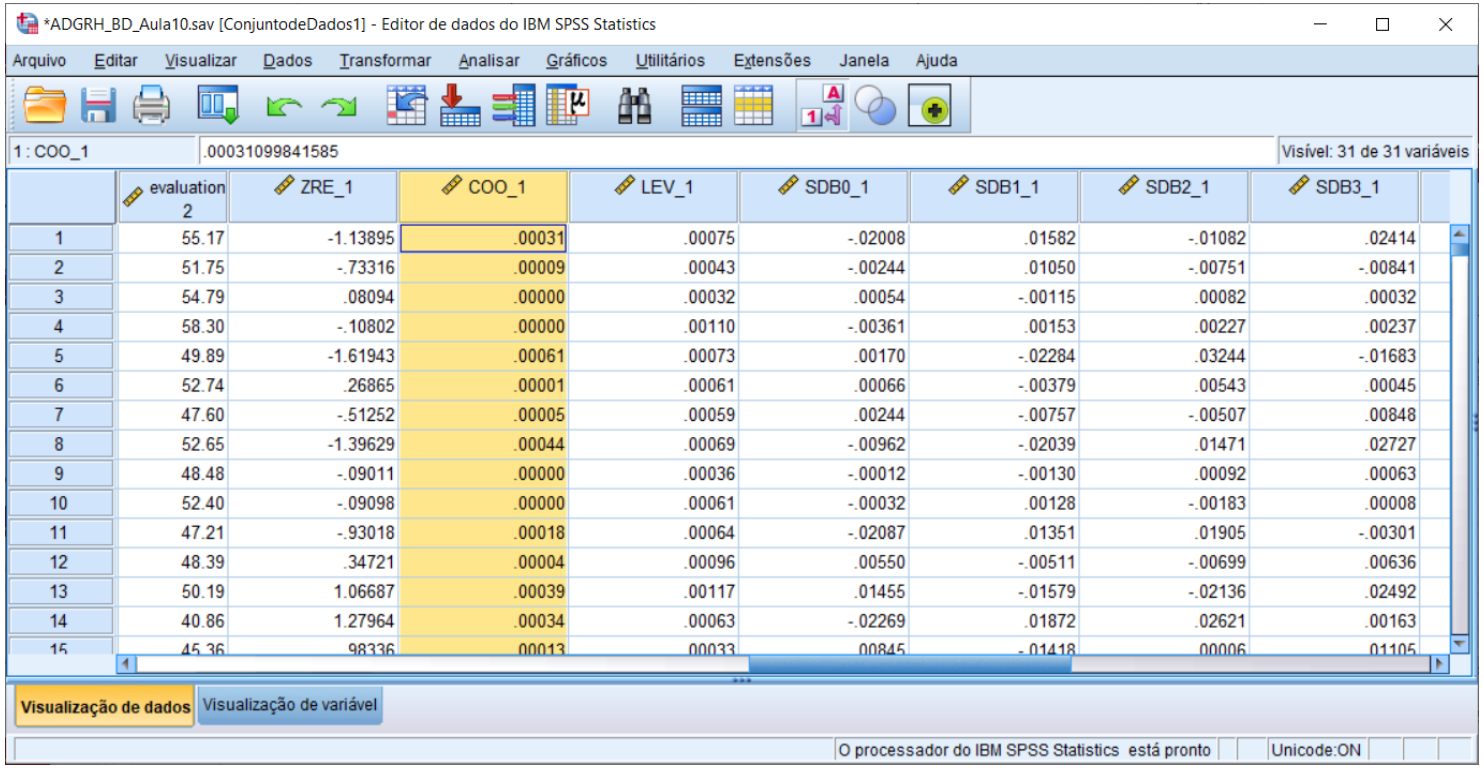
- **Quando instruímos o SPSS para produzir os diagnósticos, é criado um conjunto variáveis**
  - **Uma variável com os 'Resíduos Padronizados' da variável dependente (ZRE_1) para cada observação**



	evaluation 2	ZRE_1	LEV_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1	var
1	55.17		5	-.02008	.01582	-.01082	.02414	
2	51.75		3	-.00244	.01050	-.00751	-.00841	
3	54.79		2	.00054	-.00115	.00082	.00032	
4	58.30	-.10802	.00110	-.00361	.00153	.00227	.00237	
5	49.89	-1.61943	.00073	.00170	-.02284	.03244	-.01683	
6	52.74	.26865	.00061	.00066	-.00379	.00543	.00045	
7	47.60	-.51252	.00059	.00244	-.00757	-.00507	.00848	
8	52.65	-1.39629	.00069	-.00962	-.02039	.01471	.02727	
9	48.48	-.09011	.00036	-.00012	-.00130	.00092	.00063	
10	52.40	-.09098	.00061	-.00032	.00128	-.00183	.00008	
11	47.21	-.93018	.00064	-.02087	.01351	.01905	-.00301	
12	48.39	.34721	.00096	.00550	-.00511	-.00699	.00636	
13	50.19	1.06687	.00117	.01455	-.01579	-.02136	.02492	
14	40.86	1.27964	.00063	-.02269	.01872	.02621	.00163	
15	45.36	.98336	.00033	.00845	-.01418	.00006	.01105	

Diagnósticos

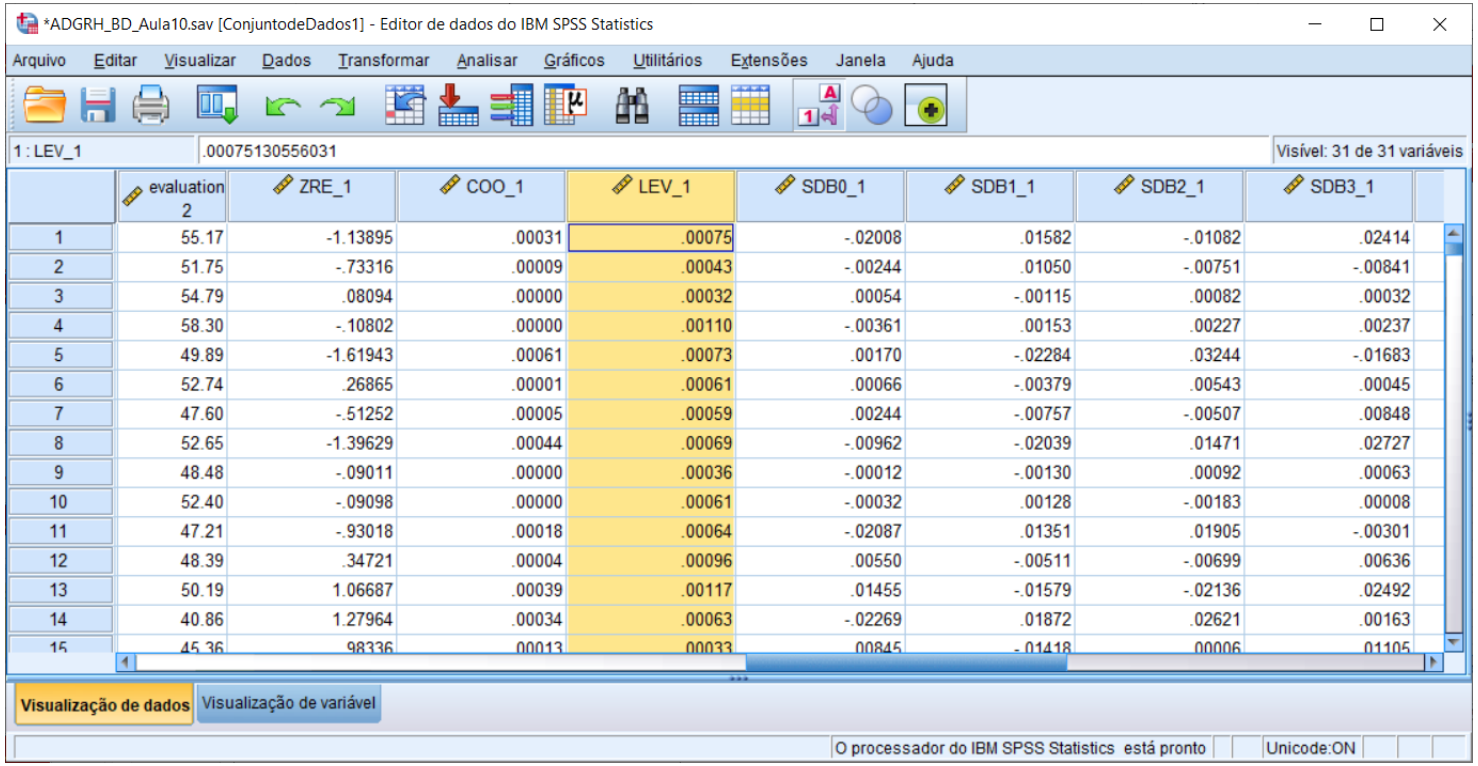
- **Quando instruímos o SPSS para produzir os diagnósticos, é criado um conjunto variáveis**
  - **Uma variável com os 'Resíduos Padronizados' da variável dependente (ZRE_1) para cada observação**
  - **Uma variável que mede a distancia de Cook associada a cada observação (COO_1)**



	evaluation 2	ZRE_1	COO_1	LEV_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1
1	55.17	-1.13895	.00031	.00075	-.02008	.01582	-.01082	.02414
2	51.75	-.73316	.00009	.00043	-.00244	.01050	-.00751	-.00841
3	54.79	.08094	.00000	.00032	.00054	-.00115	.00082	.00032
4	58.30	-.10802	.00000	.00110	-.00361	.00153	.00227	.00237
5	49.89	-1.61943	.00061	.00073	.00170	-.02284	.03244	-.01683
6	52.74	.26865	.00001	.00061	.00066	-.00379	.00543	.00045
7	47.60	-.51252	.00005	.00059	.00244	-.00757	-.00507	.00848
8	52.65	-1.39629	.00044	.00069	-.00962	-.02039	.01471	.02727
9	48.48	-.09011	.00000	.00036	-.00012	-.00130	.00092	.00063
10	52.40	-.09098	.00000	.00061	-.00032	.00128	-.00183	.00008
11	47.21	-.93018	.00018	.00064	-.02087	.01351	.01905	-.00301
12	48.39	.34721	.00004	.00096	.00550	-.00511	-.00699	.00636
13	50.19	1.06687	.00039	.00117	.01455	-.01579	-.02136	.02492
14	40.86	1.27964	.00034	.00063	-.02269	.01872	.02621	.00163
15	45.36	.98336	.00013	.00033	.00845	-.01418	.00006	.01105

Diagnósticos

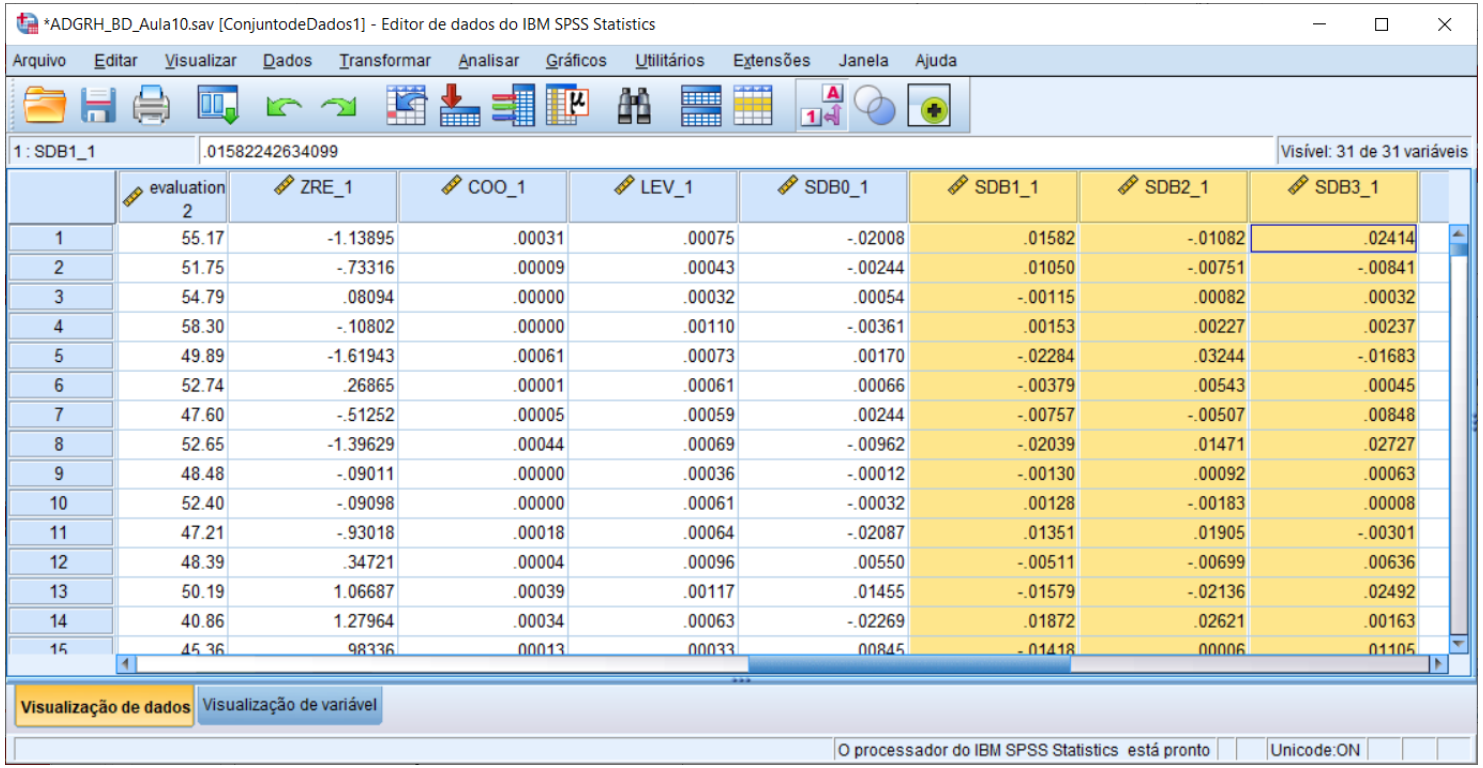
- **Quando instruímos o SPSS para produzir os diagnósticos, é criado um conjunto variáveis**
  - **Uma variável com os 'Resíduos Padronizados' da variável dependente (ZRE_1) para cada observação**
  - **Uma variável que mede a distancia de Cook associada a cada observação (COO_1)**
  - **Uma variável que mede influência relativa de cada observação no ajuste do modelo (LEV_1).**



	evaluation 2	ZRE_1	COO_1	LEV_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1
1	55.17	-1.13895	.00031	.00075	-.02008	.01582	-.01082	.02414
2	51.75	-.73316	.00009	.00043	-.00244	.01050	-.00751	-.00841
3	54.79	.08094	.00000	.00032	.00054	-.00115	.00082	.00032
4	58.30	-.10802	.00000	.00110	-.00361	.00153	.00227	.00237
5	49.89	-1.61943	.00061	.00073	.00170	-.02284	.03244	-.01683
6	52.74	.26865	.00001	.00061	.00066	-.00379	.00543	.00045
7	47.60	-.51252	.00005	.00059	.00244	-.00757	-.00507	.00848
8	52.65	-1.39629	.00044	.00069	-.00962	-.02039	.01471	.02727
9	48.48	-.09011	.00000	.00036	-.00012	-.00130	.00092	.00063
10	52.40	-.09098	.00000	.00061	-.00032	.00128	-.00183	.00008
11	47.21	-.93018	.00018	.00064	-.02087	.01351	.01905	-.00301
12	48.39	.34721	.00004	.00096	.00550	-.00511	-.00699	.00636
13	50.19	1.06687	.00039	.00117	.01455	-.01579	-.02136	.02492
14	40.86	1.27964	.00034	.00063	-.02269	.01872	.02621	.00163
15	45.36	.98336	.00013	.00033	.00845	-.01418	.00006	.01105

Diagnósticos

- **Quando instruímos o SPSS para produzir os diagnósticos, é criado um conjunto variáveis**
  - **Uma variável com os 'Resíduos Padronizados' da variável dependente (ZRE_1) para cada observação**
  - **Uma variável que mede a distancia de Cook associada a cada observação (COO_1)**
  - **Uma variável que mede influência relativa de cada observação no ajuste do modelo (LEV_1).**
  - **Por cada variável independente é criada uma variável com os DFBETA Padronizado, mede a influência de uma dada observação na estimação dos parâmetros.**



	evaluation 2	ZRE_1	COO_1	LEV_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1
1	55.17	-1.13895	.00031	.00075	-.02008	.01582	-.01082	.02414
2	51.75	-.73316	.00009	.00043	-.00244	.01050	-.00751	-.00841
3	54.79	.08094	.00000	.00032	.00054	-.00115	.00082	.00032
4	58.30	-.10802	.00000	.00110	-.00361	.00153	.00227	.00237
5	49.89	-1.61943	.00061	.00073	.00170	-.02284	.03244	-.01683
6	52.74	.26865	.00001	.00061	.00066	-.00379	.00543	.00045
7	47.60	-.51252	.00005	.00059	.00244	-.00757	-.00507	.00848
8	52.65	-1.39629	.00044	.00069	-.00962	-.02039	.01471	.02727
9	48.48	-.09011	.00000	.00036	-.00012	-.00130	.00092	.00063
10	52.40	-.09098	.00000	.00061	-.00032	.00128	-.00183	.00008
11	47.21	-.93018	.00018	.00064	-.02087	.01351	.01905	-.00301
12	48.39	.34721	.00004	.00096	.00550	-.00511	-.00699	.00636
13	50.19	1.06687	.00039	.00117	.01455	-.01579	-.02136	.02492
14	40.86	1.27964	.00034	.00063	-.02269	.01872	.02621	.00163
15	45.36	.98336	.00013	.00033	.00845	-.01418	.00006	.01105

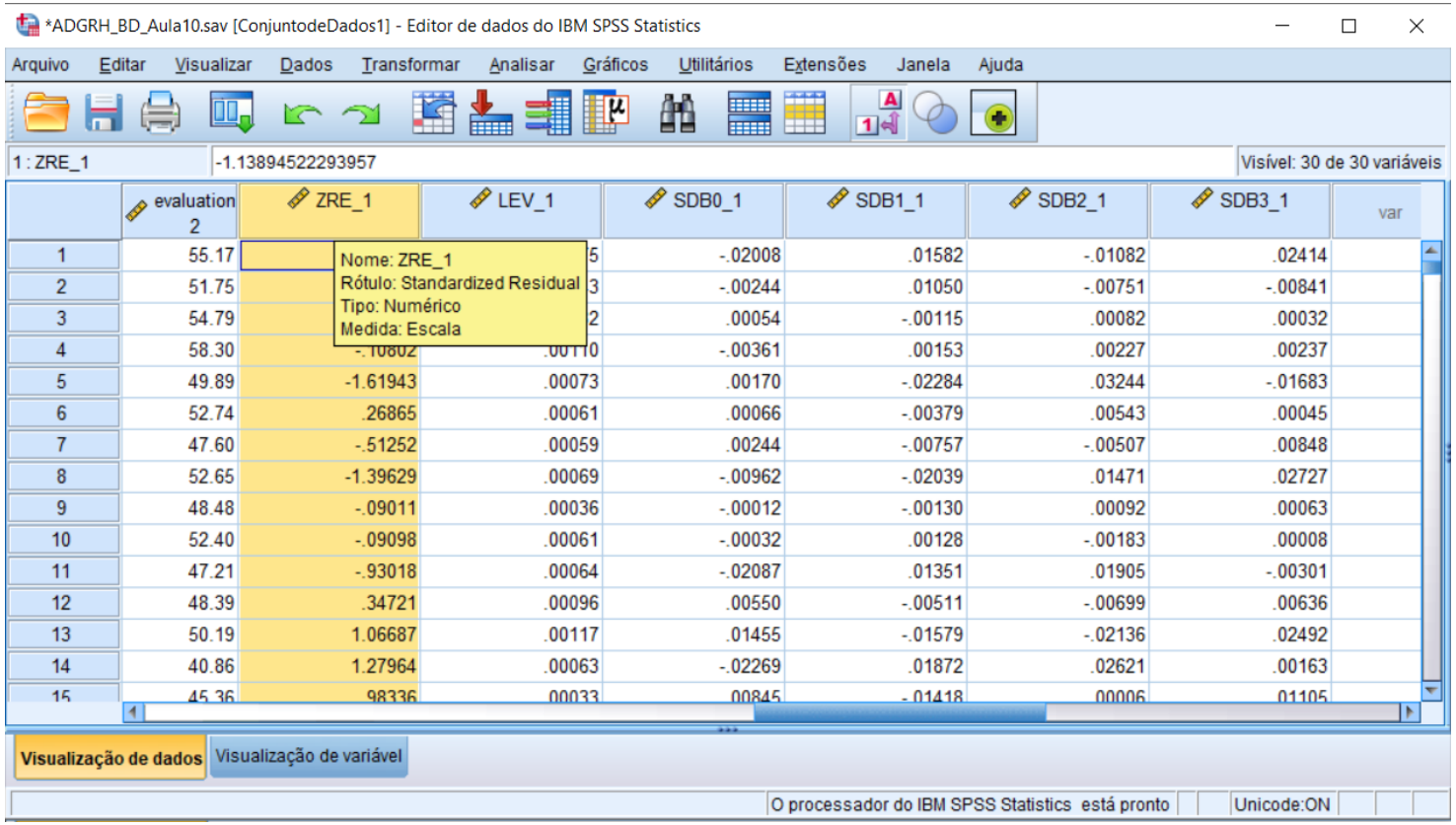
Validação do Modelo de Regressão Linear

2. Avaliação do Pressuposto II: Normalidade da Distribuição dos Erros

Normalidade da Distribuição dos Erros

- **Para avaliarmos se os erros seguem uma distribuição normal, vamos usar a variável com os 'Resíduos Padronizados' da VD (ZRE_1) que acabamos de criar.**

- **Vamos então criar usar um gráfico Q-Q para representar a distribuição dos resíduos padronizados**



	evaluation 2	ZRE_1	LEV_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1	var
1	55.17			-.02008	.01582	-.01082	.02414	
2	51.75			-.00244	.01050	-.00751	-.00841	
3	54.79			.00054	-.00115	.00082	.00032	
4	58.30			-.00361	.00153	.00227	.00237	
5	49.89	-1.61943	.00073	.00170	-.02284	.03244	-.01683	
6	52.74	.26865	.00061	.00066	-.00379	.00543	.00045	
7	47.60	-.51252	.00059	.00244	-.00757	-.00507	.00848	
8	52.65	-1.39629	.00069	-.00962	-.02039	.01471	.02727	
9	48.48	-.09011	.00036	-.00012	-.00130	.00092	.00063	
10	52.40	-.09098	.00061	-.00032	.00128	-.00183	.00008	
11	47.21	-.93018	.00064	-.02087	.01351	.01905	-.00301	
12	48.39	.34721	.00096	.00550	-.00511	-.00699	.00636	
13	50.19	1.06687	.00117	.01455	-.01579	-.02136	.02492	
14	40.86	1.27964	.00063	-.02269	.01872	.02621	.00163	
15	45.36	.98336	.00033	.00845	-.01418	.00006	.01105	

Normalidade da Distribuição dos Erros

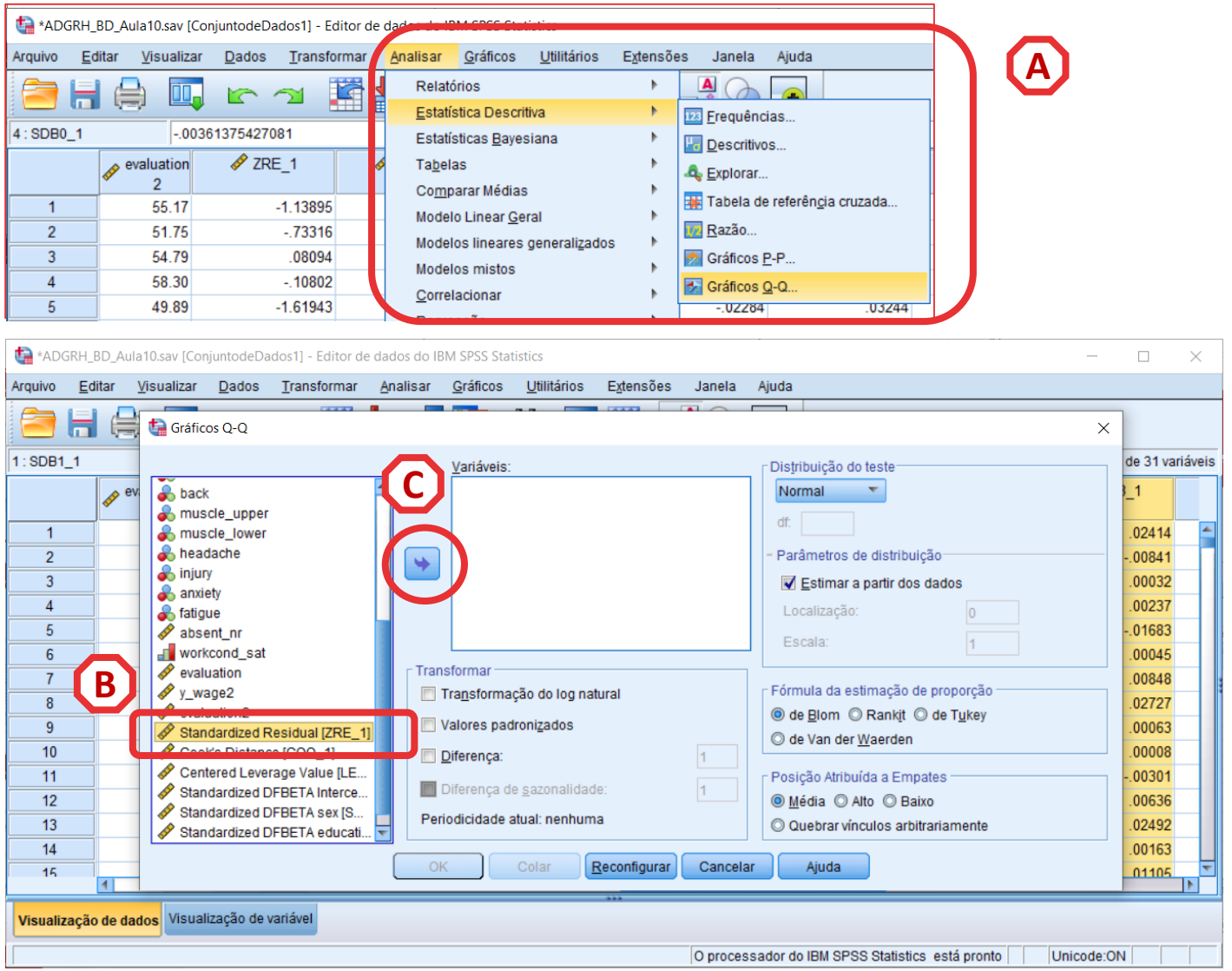
- **Selecionar ‘Analisar’ / ‘Estatística Descritiva’ / ‘Gráficos Q-Q’**
- **Selecionar a variável ‘ZRE_1’**
- **Colocar na caixa ‘Variáveis’**
- **Selecionar ‘OK’**

A

B

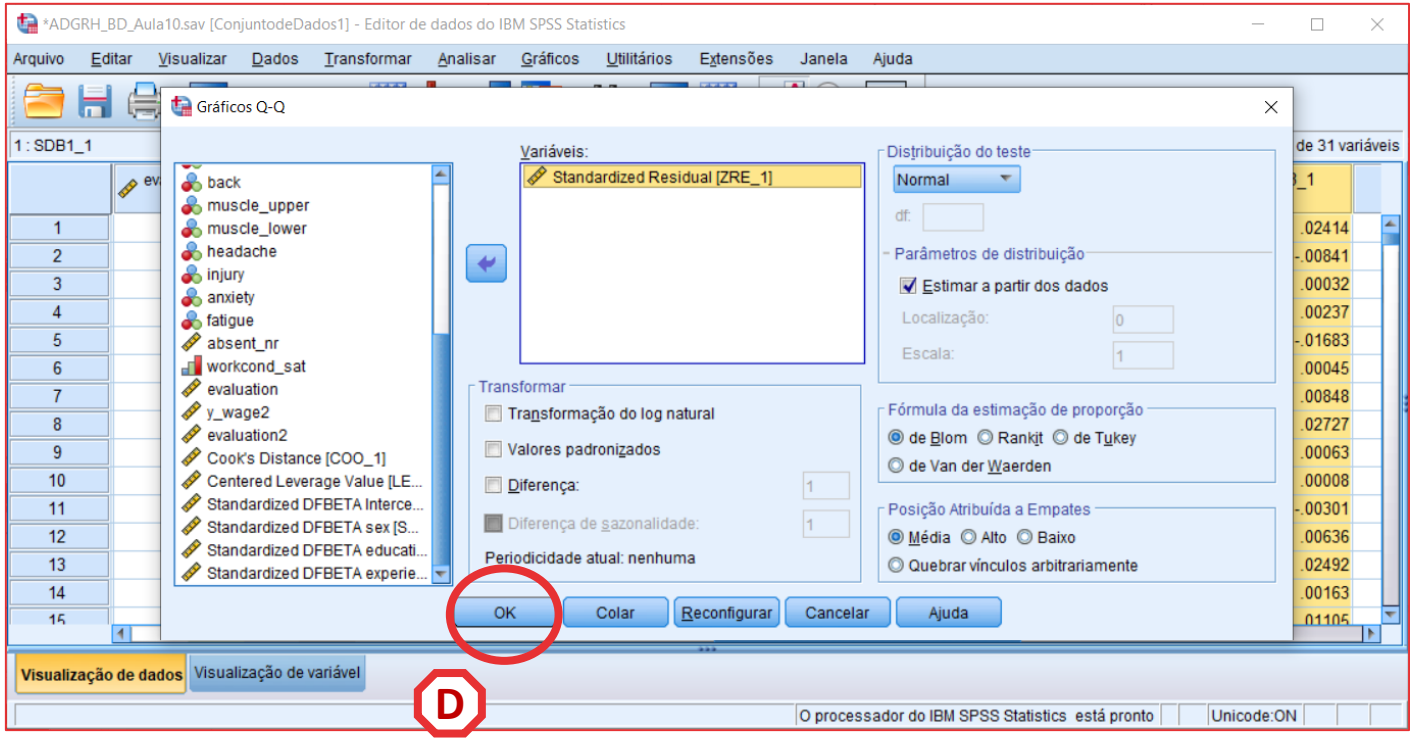
C

D



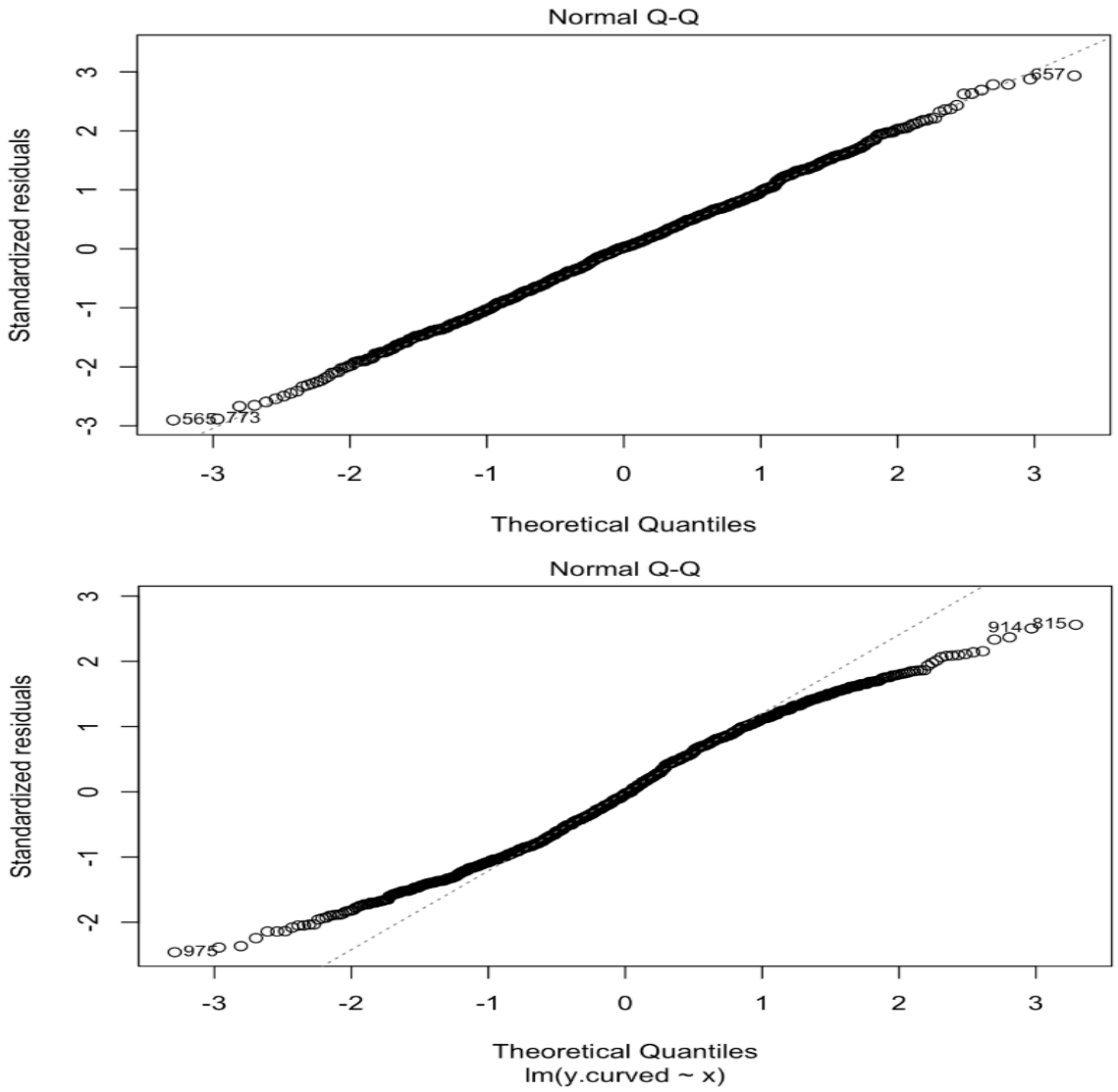
Normalidade da Distribuição dos Erros

- Selecionar ‘Analisar’ / ‘Estatística Descritiva’ / ‘Gráficos Q-Q’
- Selecionar a variável ‘ZRE_1’
- Colocar na caixa ‘Variáveis’
- Selecionar ‘OK’



Normalidade da Distribuição dos Erros

- **Linha diagonal reflecte uma distribuição normal**
- **Os resíduos sobrepõe-se quase totalmente com a linha de diagonal**
- **Os resíduos parecem estar normalmente distribuídos**
- **Neste, caso os as caudas da distribuição dos resíduos afasta-se da diagonal, o que sugere que a distribuição dos erros pode não ser normal**

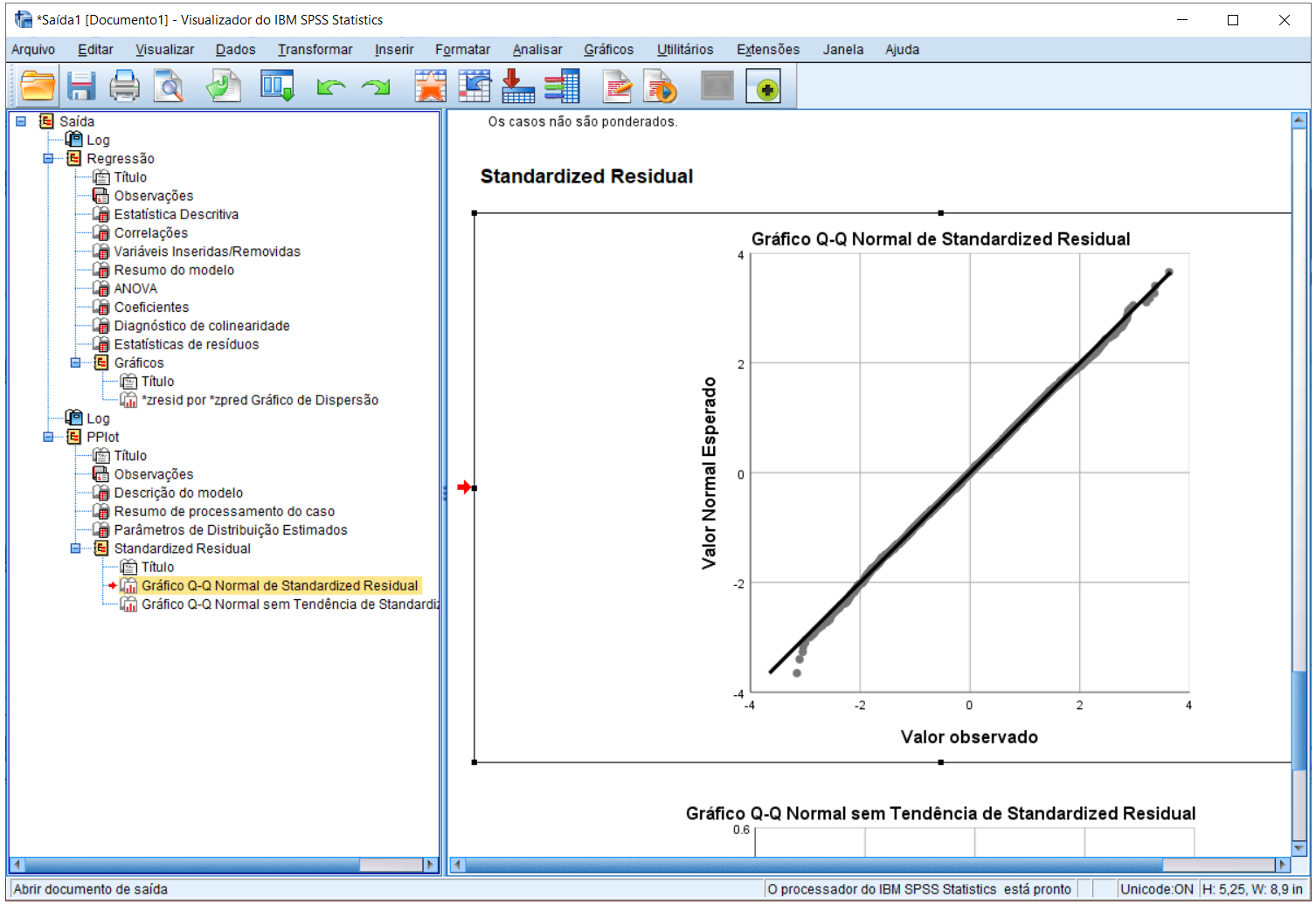


Fonte: https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html

Normalidade da Distribuição dos Erros

- **O gráfico é publicado no 'Visualizador de Resultados'**

- **Neste caso podemos concluir que os erros seguem uma distribuição normal!**



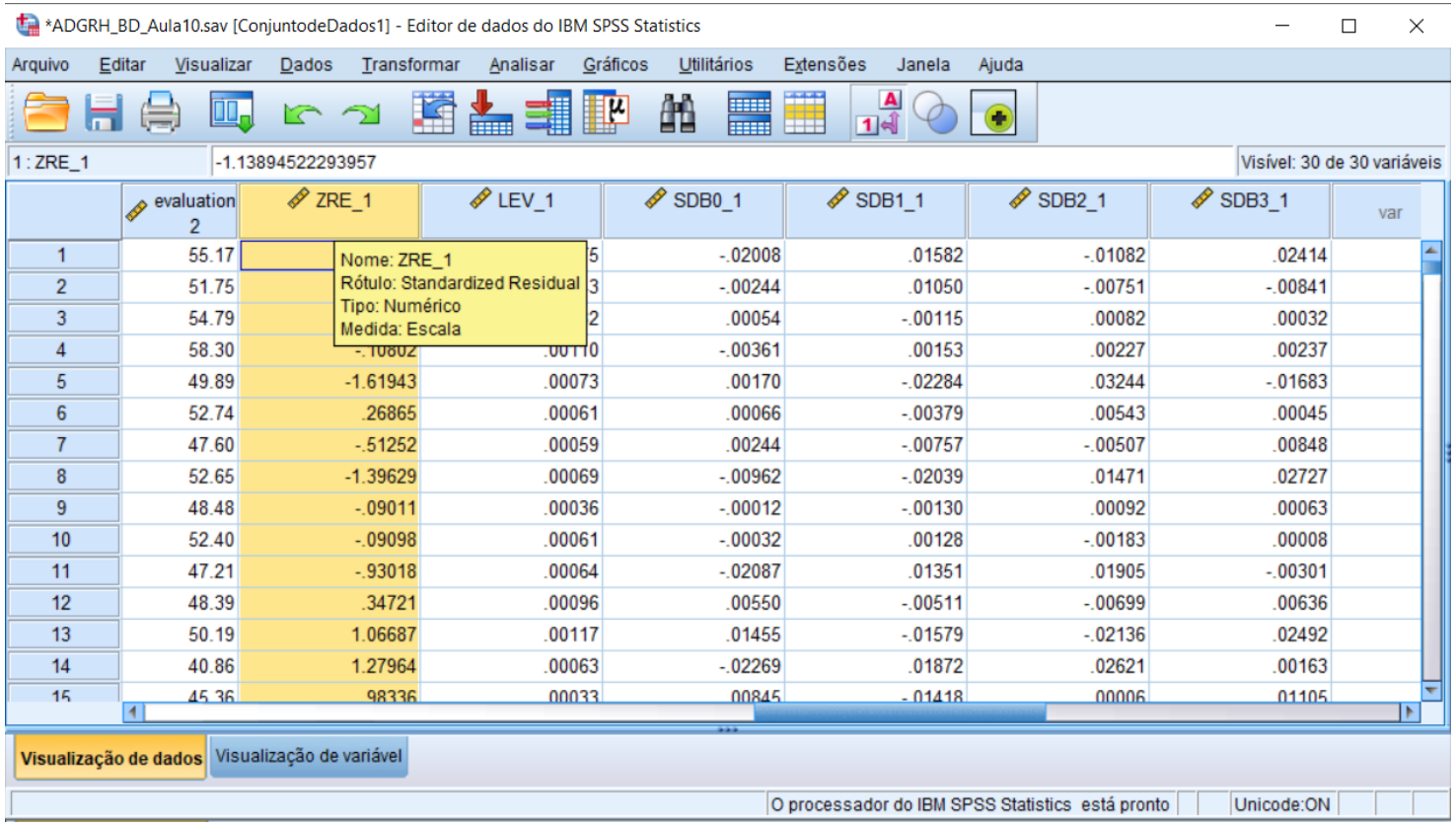
Validação do Modelo de Regressão Linear

3. Avaliação do Pressuposto III: Média Condicional Zero dos Erros

Média Condicional Zero dos Erros

- **Para avaliarmos se o termo de erro aleatório tem valor esperado igual a zero, vamos usar a variável com os 'Resíduos Padronizados' da VD (ZRE_1) que acabamos de criar.**

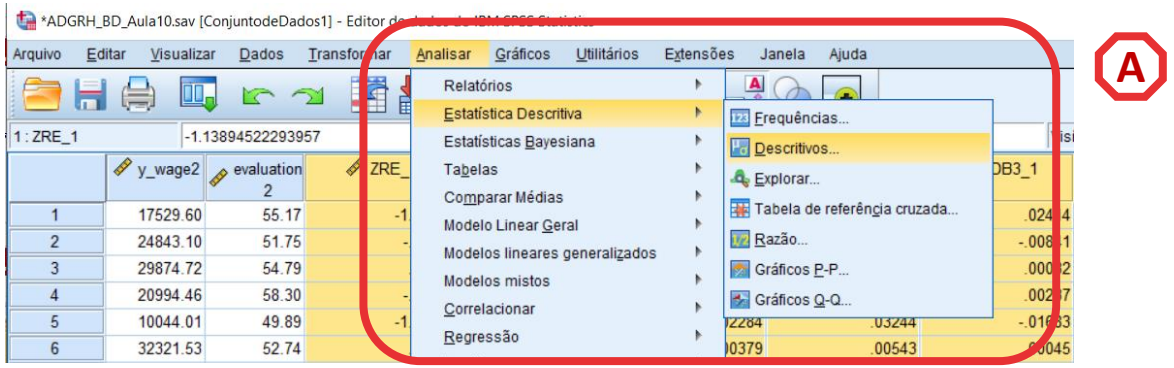
- **Mas neste caso, vamos olhar para as estatísticas descritivas desta variável.**

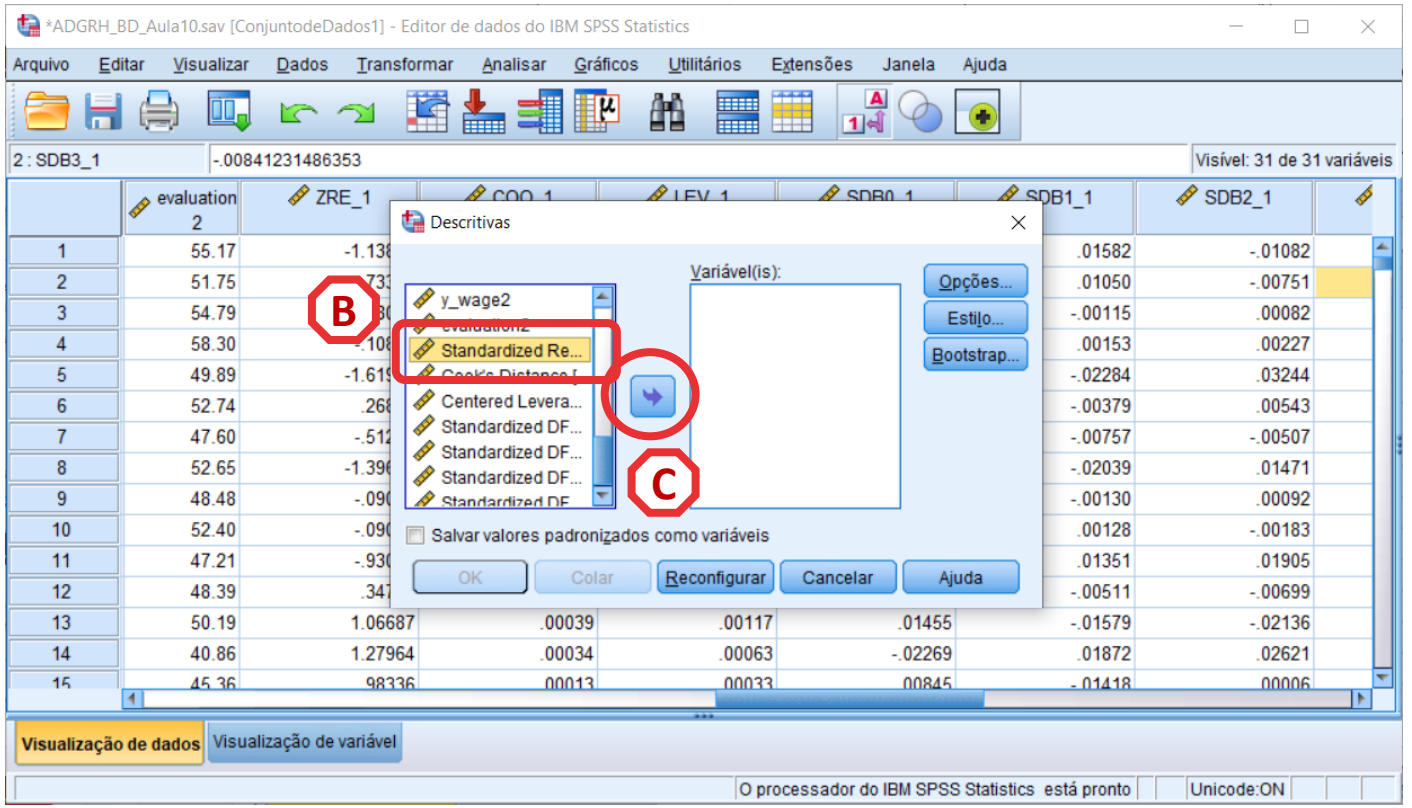


	evaluation 2	ZRE_1	LEV_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1	var
1	55.17		5	-.02008	.01582	-.01082	.02414	
2	51.75		3	-.00244	.01050	-.00751	-.00841	
3	54.79		2	.00054	-.00115	.00082	.00032	
4	58.30	-.10802	.00110	-.00361	.00153	.00227	.00237	
5	49.89	-1.61943	.00073	.00170	-.02284	.03244	-.01683	
6	52.74	.26865	.00061	.00066	-.00379	.00543	.00045	
7	47.60	-.51252	.00059	.00244	-.00757	-.00507	.00848	
8	52.65	-1.39629	.00069	-.00962	-.02039	.01471	.02727	
9	48.48	-.09011	.00036	-.00012	-.00130	.00092	.00063	
10	52.40	-.09098	.00061	-.00032	.00128	-.00183	.00008	
11	47.21	-.93018	.00064	-.02087	.01351	.01905	-.00301	
12	48.39	.34721	.00096	.00550	-.00511	-.00699	.00636	
13	50.19	1.06687	.00117	.01455	-.01579	-.02136	.02492	
14	40.86	1.27964	.00063	-.02269	.01872	.02621	.00163	
15	45.36	.98336	.00033	.00845	-.01418	.00006	.01105	

Média Condicional Zero dos Erros

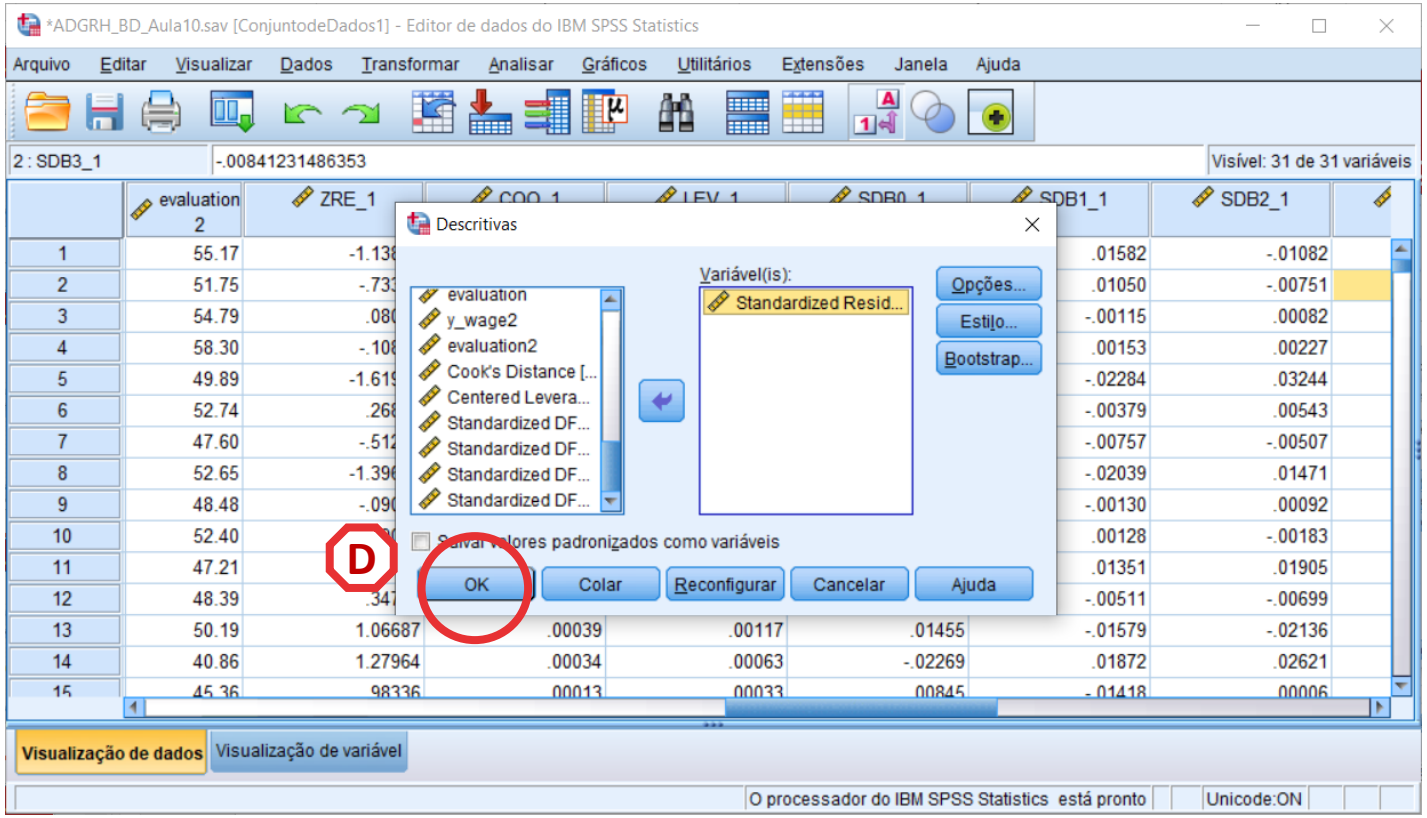
- Selecionar ‘Analisar’ / ‘Estatística Descritiva’ / ‘Descritivos’ (A)
- Selecionar a variável ‘ZRE_1’ (B)
- Colocar na caixa ‘Variável(is)’ (C)
- Selecionar ‘OK’ (D)





Média Condicional Zero dos Erros

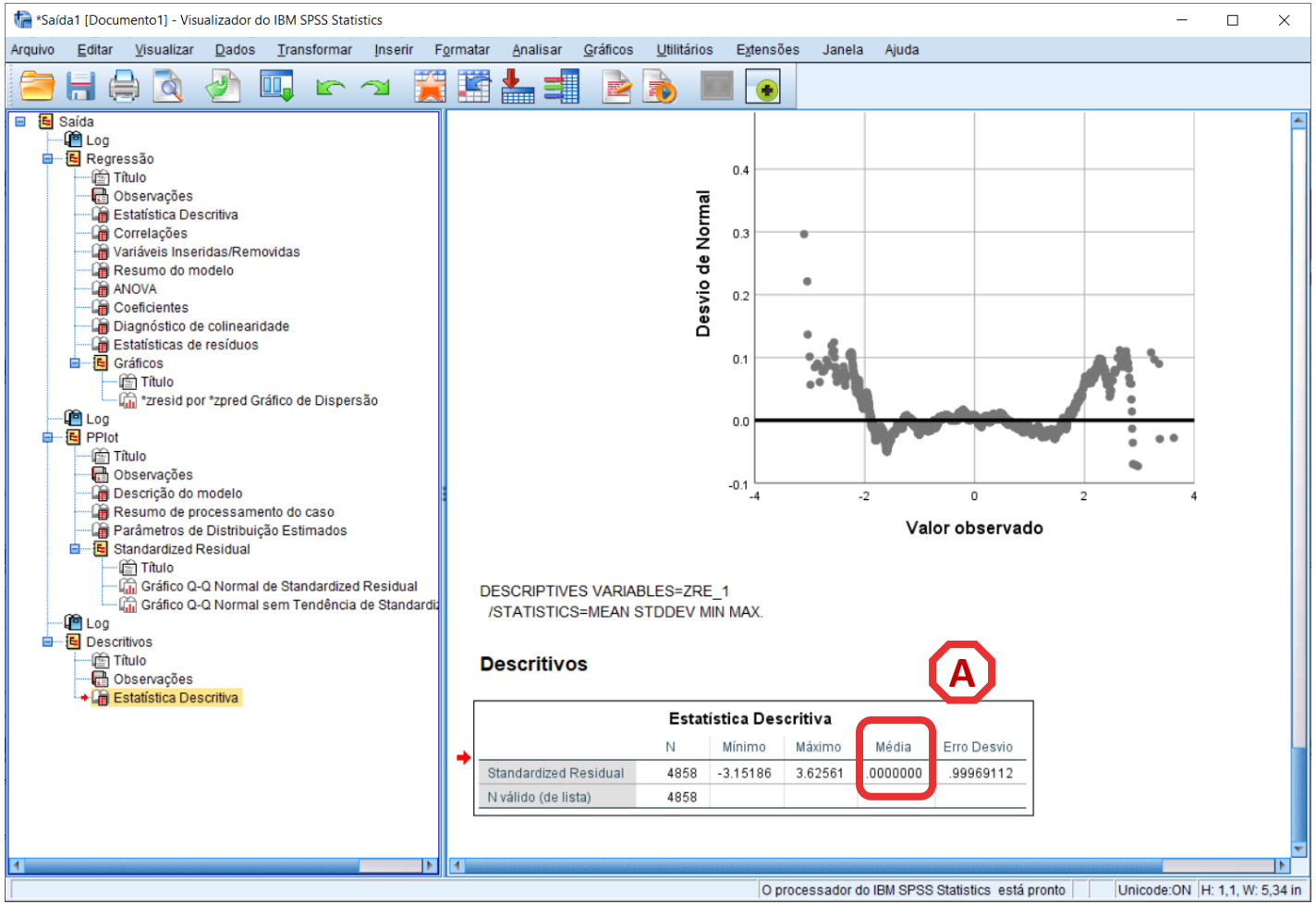
- Selecionar ‘Analisar’ / ‘Estatística Descritiva’ / ‘Descritivos’ (A)
- Selecionar a variável ‘ZRE_1’ (B)
- Colocar na caixa ‘Variável(is)’ (C)
- Selecionar ‘OK’ (D)



Média Condicional Zero dos Erros

- O gráfico é publicado no ‘Visualizador de Resultados’
- Os ‘Resíduos Padronizados’ da VD (ZRE_1) tem uma média muito próximo de 0,
- Neste caso podemos concluir que se cumpre o pressuposto da Média Condicional Zero dos Erros.

A



Validação do Modelo de Regressão Linear

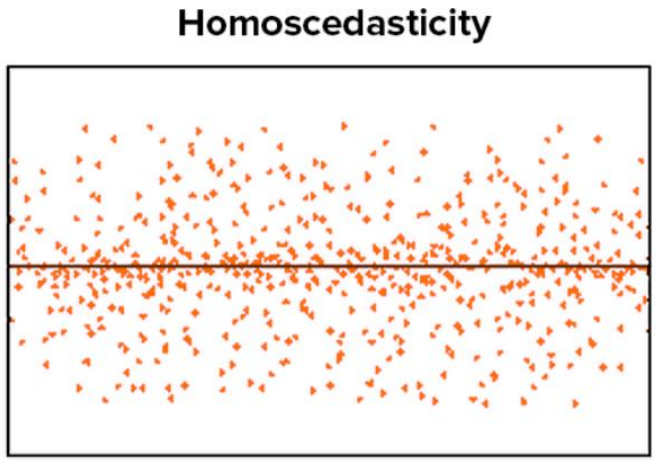
4. Avaliação do Pressuposto IV: Homocedasticidade

Homocedasticidade (ou Igual Variância)

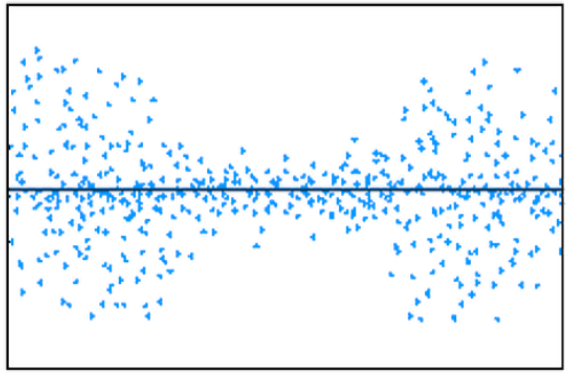
- **A distribuição dos resíduos apresenta uma variância constante ao longo dos valores previstos da variável dependente. Não há indicação de variação não-constante.**

- **Neste, o valor dos resíduos aproxima-se de 0 para os valores mais baixos da predição, mas aumentam à medida que os valores previstos também aumenta.**

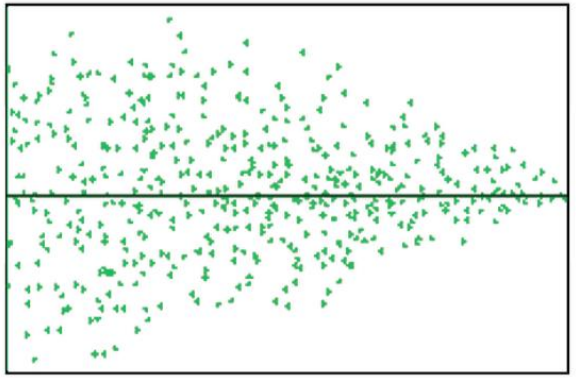
ou seja, a variação não é constante.



Random Cloud (No Discernible Pattern)



Bow Tie Shape (Pattern)



Fan Shape (Pattern)

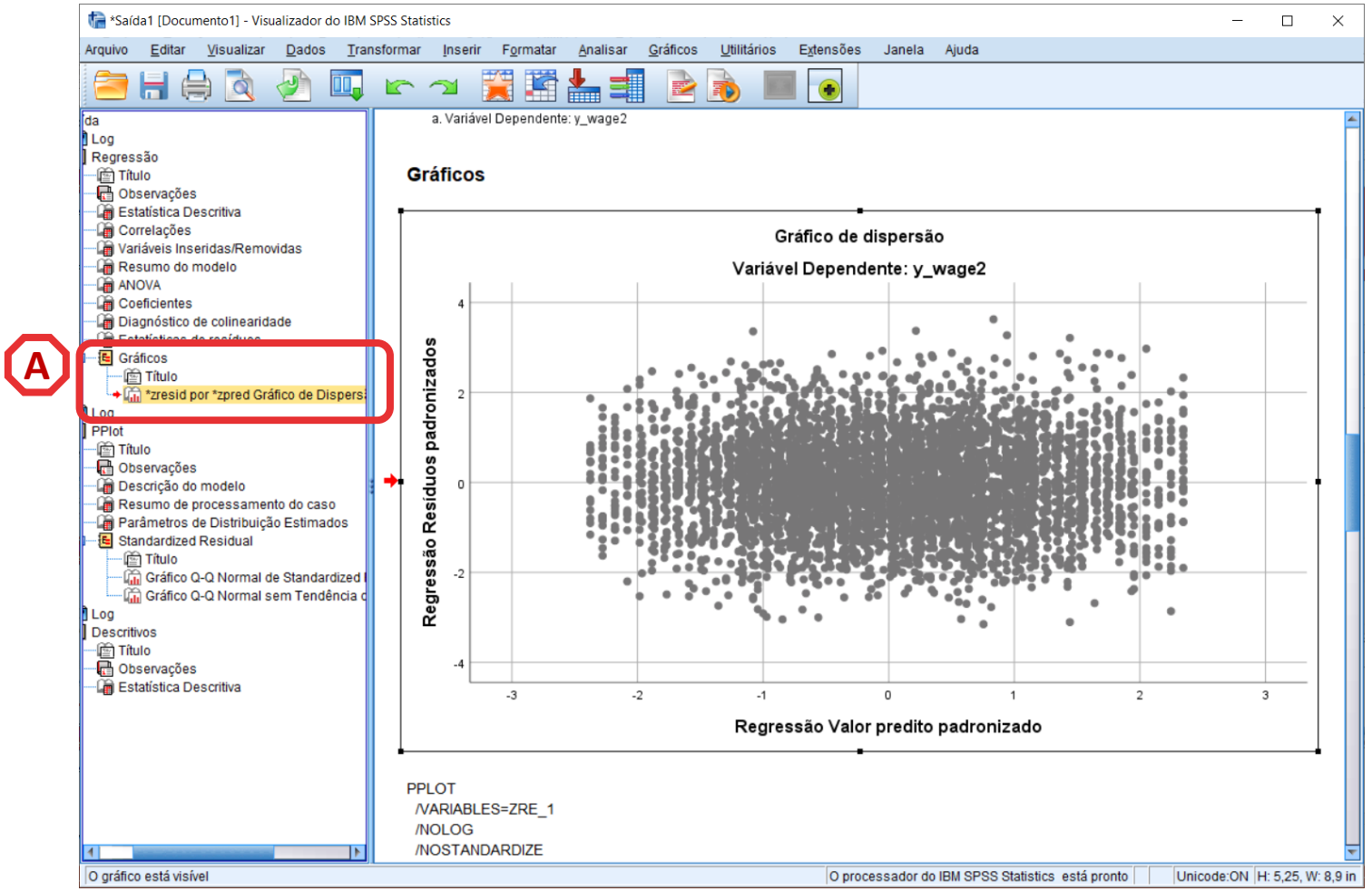
Fonte: https://towardsdatascience.com/assumptions-of-linear-regression-algorithm-ed9ea32224e1

Homocedasticidade

- Para avaliar se se cumpre este pressuposto, temos de olhar para o Gráfico de Dispersão que compara a distribuição dos 'Residuos Padronizados' com os 'Valores Preditos Padronizados' - que o SPSS produz automaticamente.

- Neste caso, a representação da distribuição parece sugerir que a variação dos resíduos é relativamente constante.

- Ou seja, cumpre-se o pressuposto da Homocedasticidade



Validação do Modelo de Regressão Linear

5. Avaliação do Pressuposto V: Independência dos Erros

Independência dos Erros

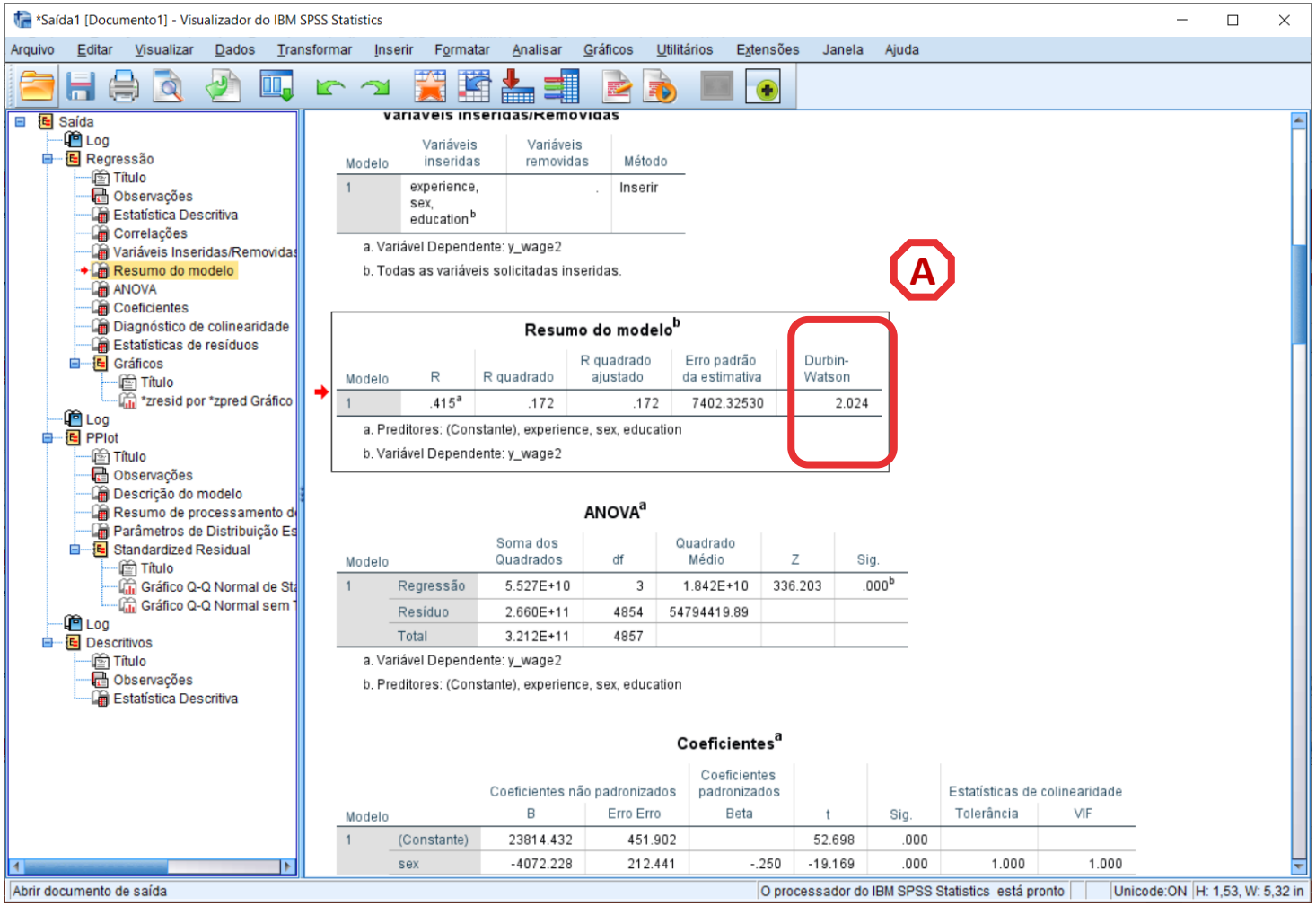
- **Para avaliar se se cumpre este pressuposto, temos de olhar para o resultado do teste Durbin-Watson - que pedimos ao SPSS para produzir.**

- **Interpretação:**

= 2 -> Erros <u>são</u> independentes

> 2 / <2 -> Erros <u>não são</u> independentes

- **Neste caso os erros são independentes**



Validação do Modelo de Regressão Linear

6. Avaliação do Pressuposto VI: Ausência de Multicolinearidade Perfeita

Ausência de Multicolienaridade Perfeita

- **Quando há fortes relações lineares entre os preditores numa regressão, a precisão dos coeficientes de regressão diminui em comparação com o que teria sido se os preditores não se correlacionassem entre si**

- **Um valor de VIF > 3 sugere a existência de colinearidade no modelo**
- **Um valor de VIF > 10 sugere a existência de colinearidade séria**

- **Deve repensar-se as variáveis a incluir no modelo**

Ausência de Multicolinearidade

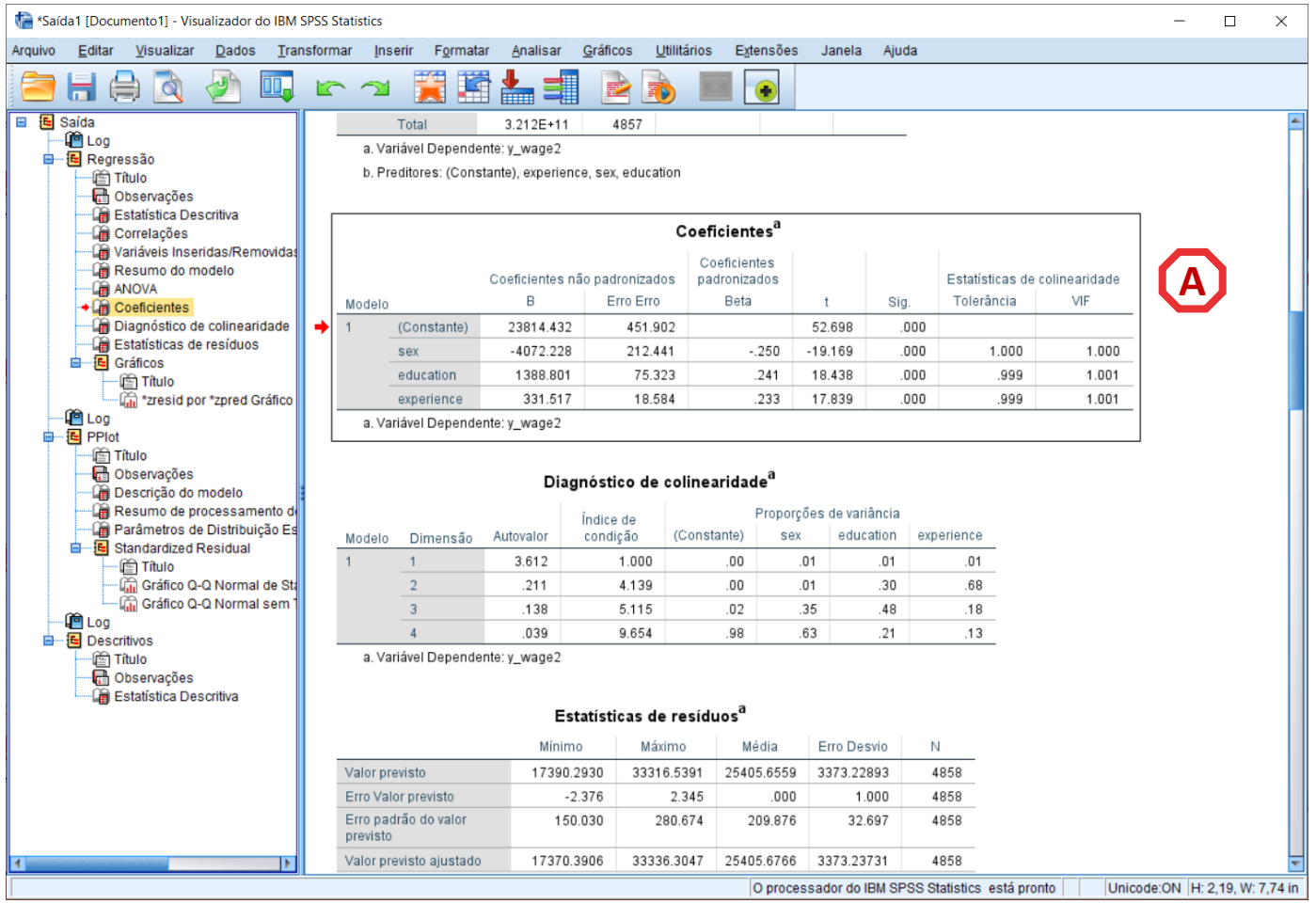
- Para testarmos este pressuposto, temos de olhar para a Tabela de Coeficientes - que o SPSS produz automaticamente.

- Interpretação

VIF > 3 -> presença de colinearidade

- Neste caso, não se identifica a presença de colinearidade…

- Portanto, cumpre-se o pressuposto da ausência de Multicolinearidade

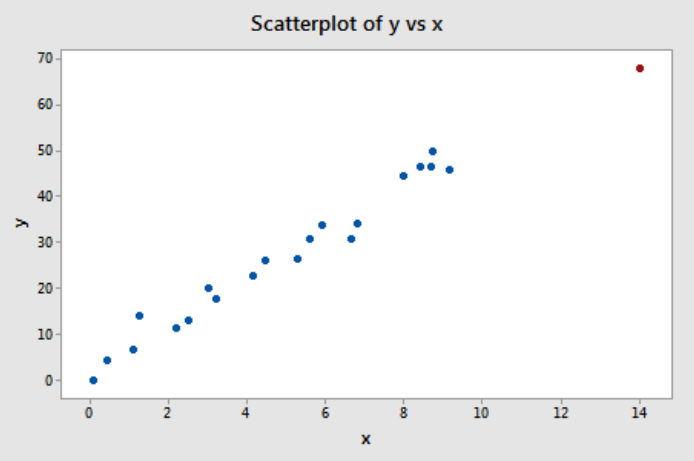


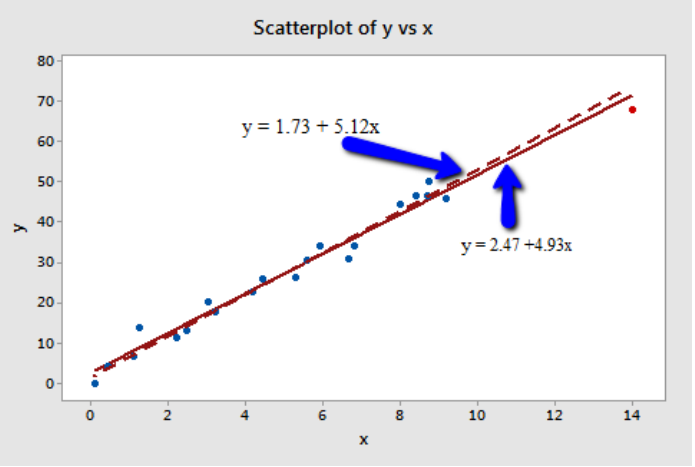
Validação do Modelo de Regressão Linear

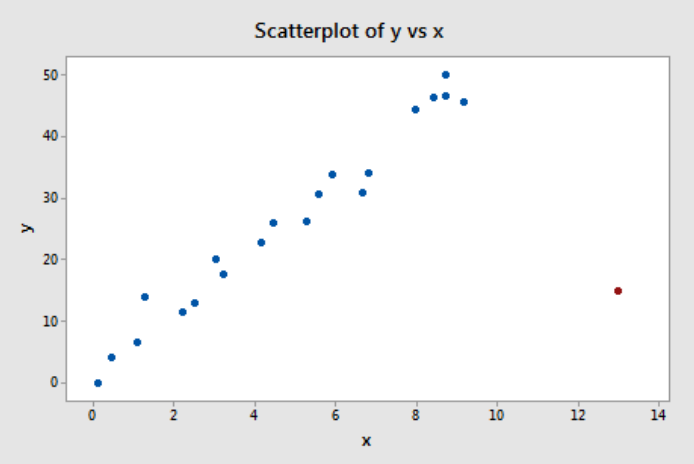
7. Avaliação do Pressuposto VII: Ausência de Observações Influentes

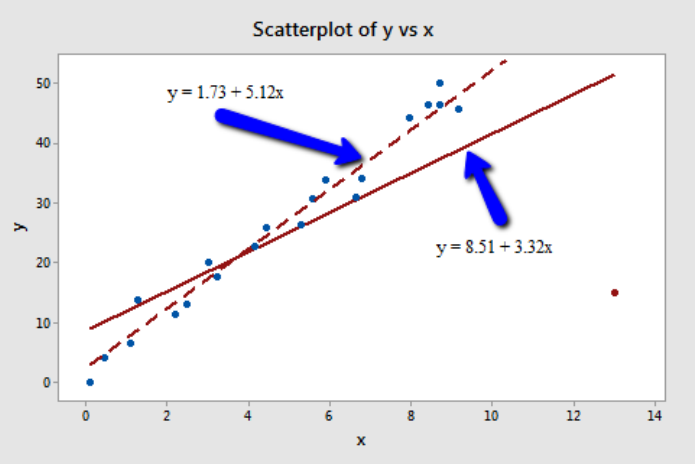
Ausência de Observações Influentes

- A existência de 'Outliers' (valores extremos) não é um problema em si.
- Mas torna-se um problema quando os Outlier têm influência sobre os resultados do modelo
- Nos painéis de baixo, o Outlier é uma 'Observação Influente'



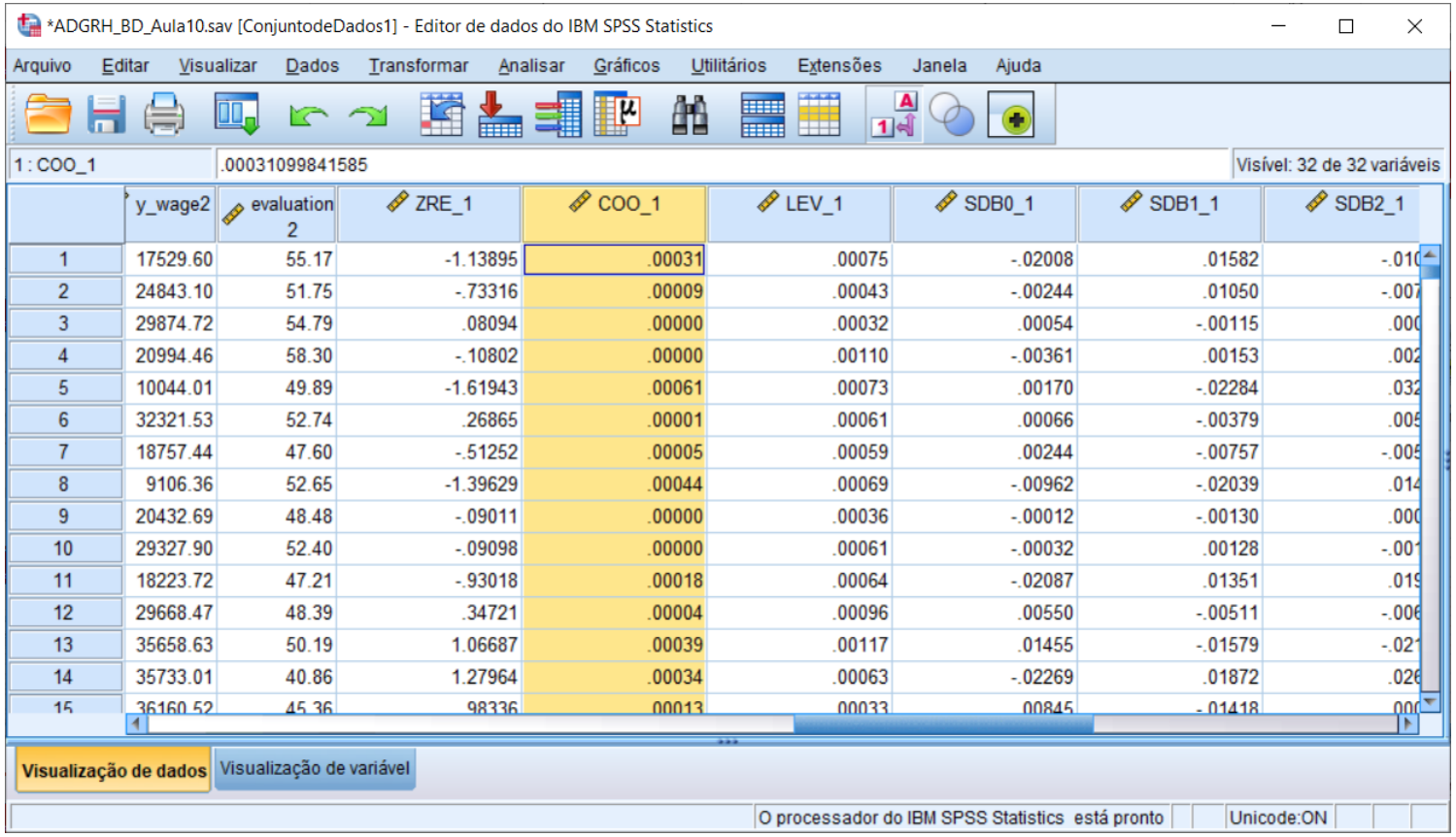






Ausência de Observações Influentes

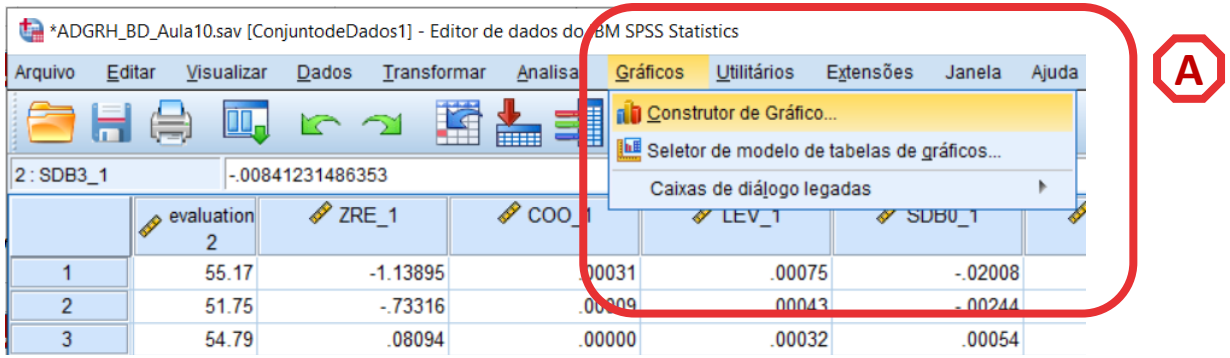
- **Para testarmos a presença de observações influentes vamos usar a variável com os 'Distância de Cook' (COO_1) que acabamos de criar.**

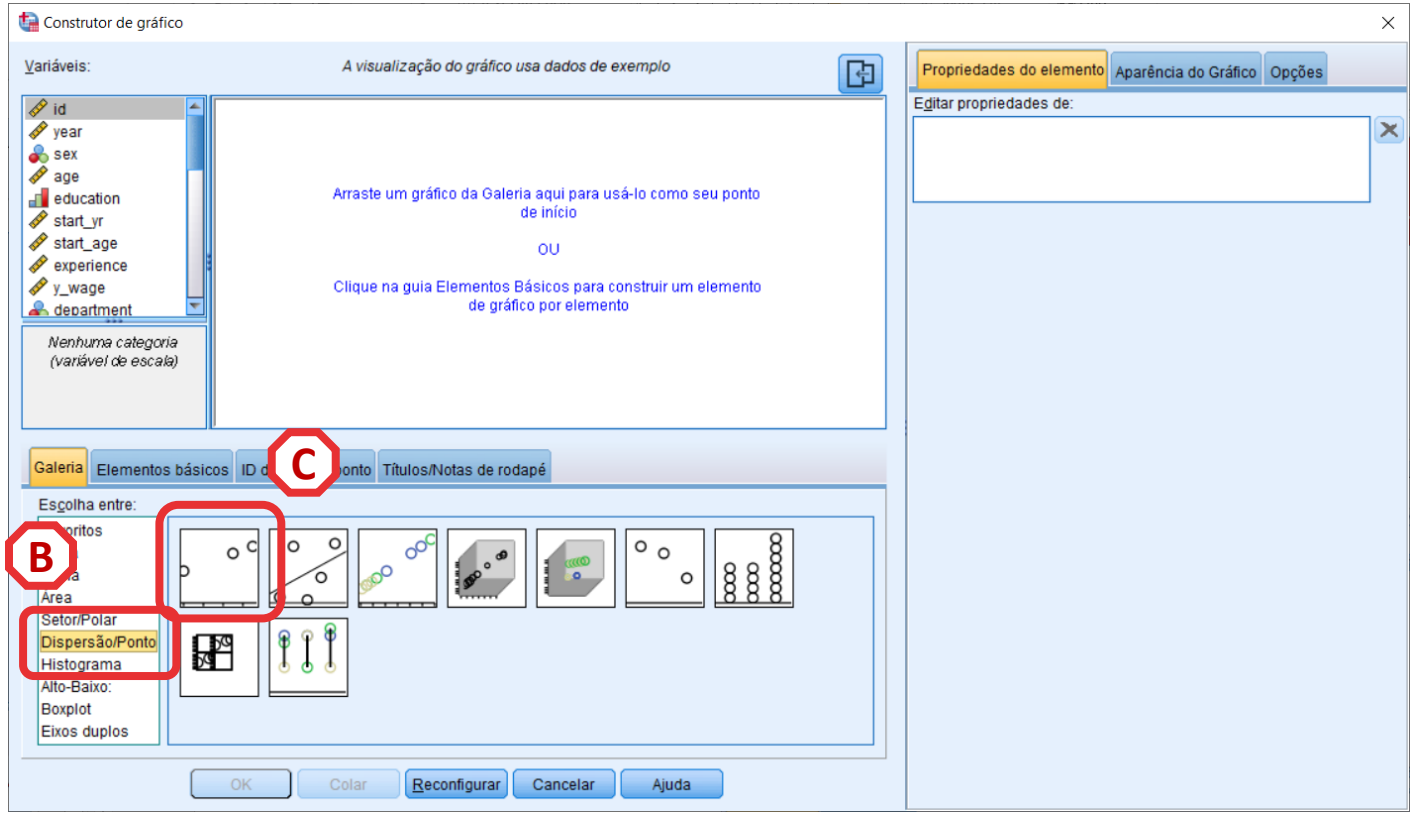


	y_wage2	evaluation 2	ZRE_1	COO_1	LEV_1	SDB0_1	SDB1_1	SDB2_1
1	17529.60	55.17	-1.13895	.00031	.00075	-.02008	.01582	-.01
2	24843.10	51.75	-.73316	.00009	.00043	-.00244	.01050	-.007
3	29874.72	54.79	.08094	.00000	.00032	.00054	-.00115	.000
4	20994.46	58.30	-.10802	.00000	.00110	-.00361	.00153	.002
5	10044.01	49.89	-1.61943	.00061	.00073	.00170	-.02284	.032
6	32321.53	52.74	.26865	.00001	.00061	.00066	-.00379	.005
7	18757.44	47.60	-.51252	.00005	.00059	.00244	-.00757	-.005
8	9106.36	52.65	-1.39629	.00044	.00069	-.00962	-.02039	.014
9	20432.69	48.48	-.09011	.00000	.00036	-.00012	-.00130	.000
10	29327.90	52.40	-.09098	.00000	.00061	-.00032	.00128	-.001
11	18223.72	47.21	-.93018	.00018	.00064	-.02087	.01351	.019
12	29668.47	48.39	.34721	.00004	.00096	.00550	-.00511	-.006
13	35658.63	50.19	1.06687	.00039	.00117	.01455	-.01579	-.021
14	35733.01	40.86	1.27964	.00034	.00063	-.02269	.01872	.026
15	36160.52	45.36	.98336	.00013	.00033	.00845	-.01418	.000

Ausência de Observações Influentes

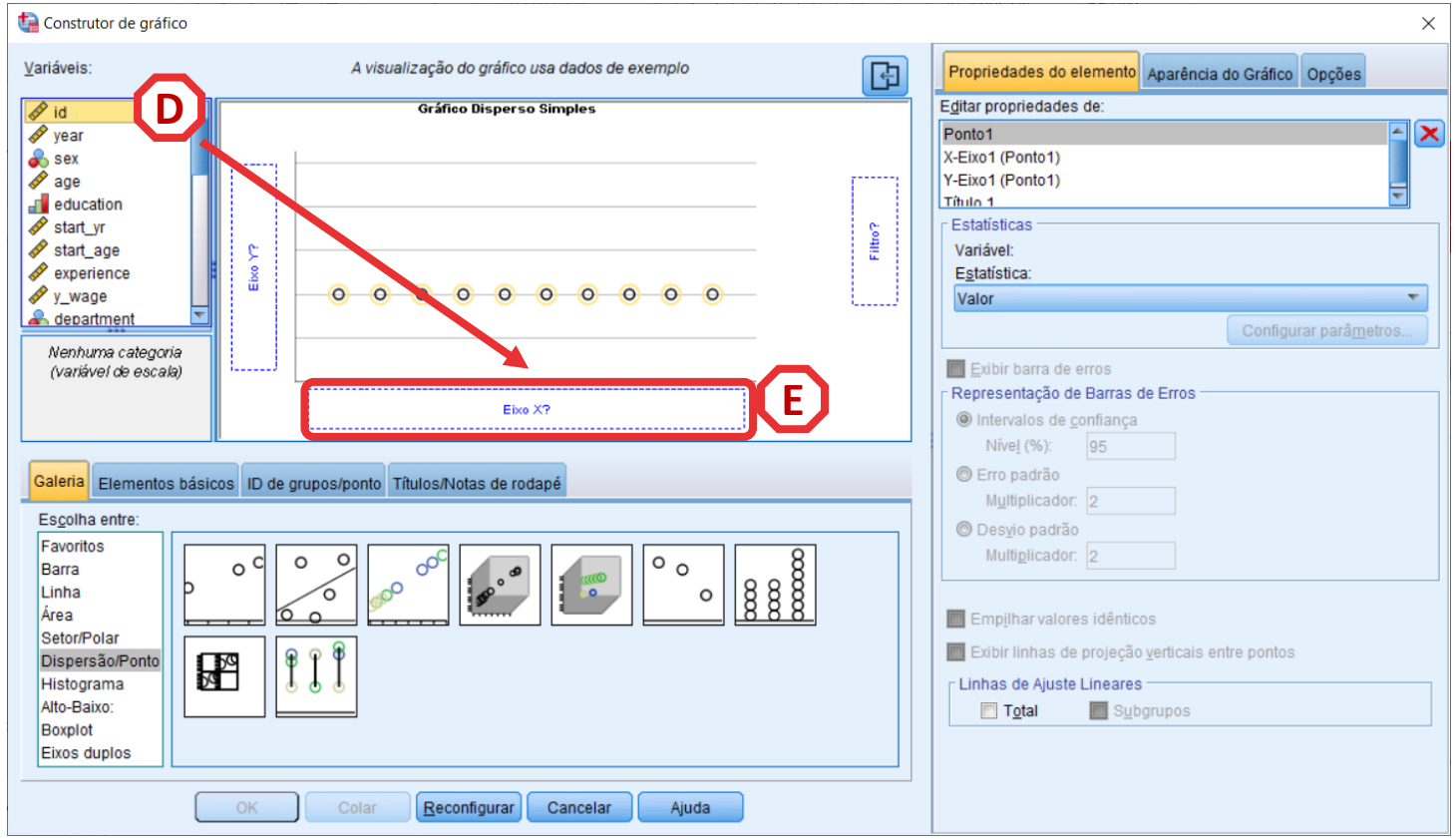
- Selecionar ‘Gráficos’ / ‘Construtor de Gráfico’ (A)
- Selecionar ‘DispersãoPontos’ (B)
- Selecionar ‘Dispersão (Simples)’ (C)





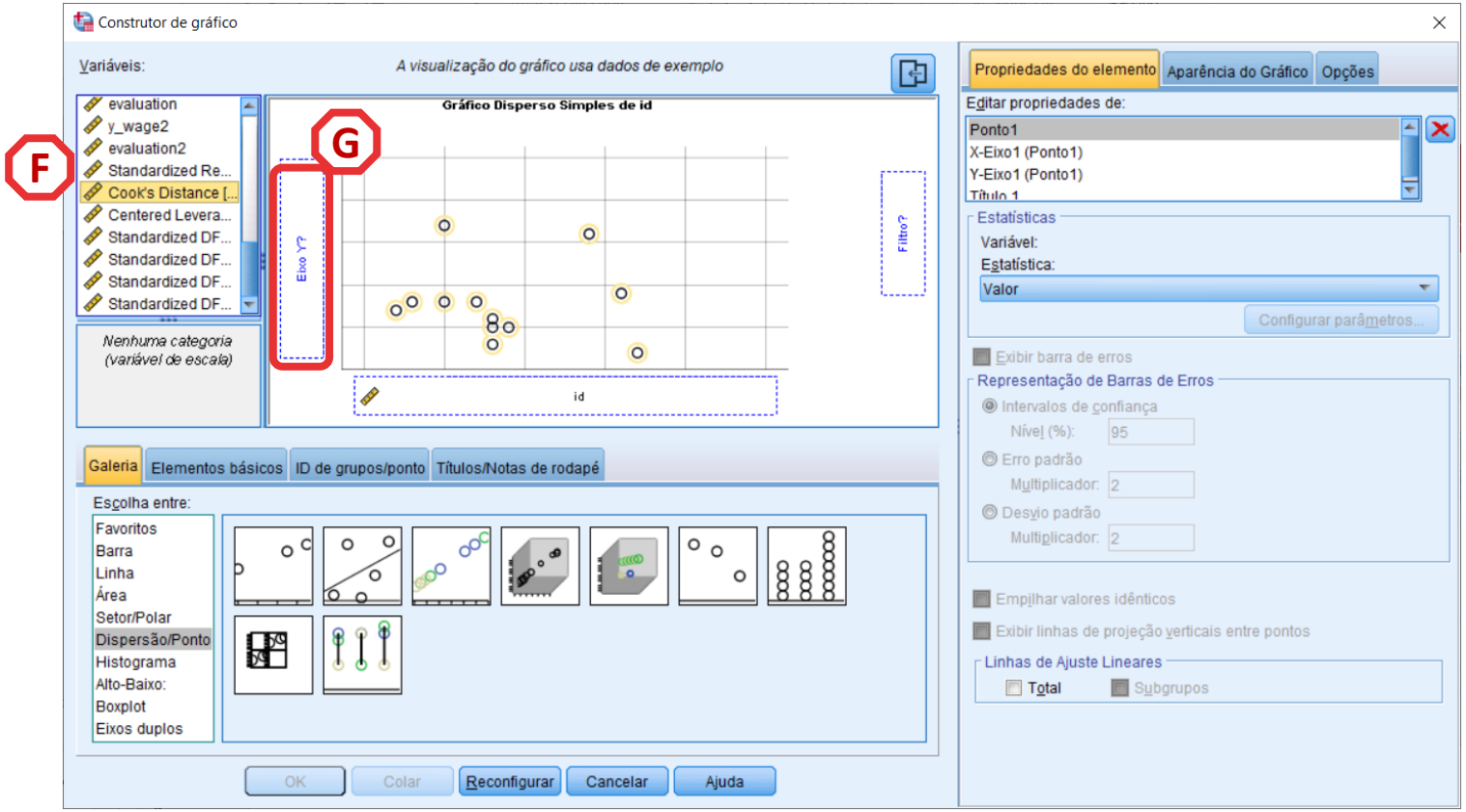
Ausência de Observações Influentes

- Selecionar ‘Gráficos’ / ‘Construtor de Gráfico’ (A)
- Selecionar ‘DispersãoPontos’ (B)
- Selecionar ‘Dispersão (Simples)’ (C)
- Selecionar Variável ‘id’ (D)
- Colocar no eixo ‘x’ (E)



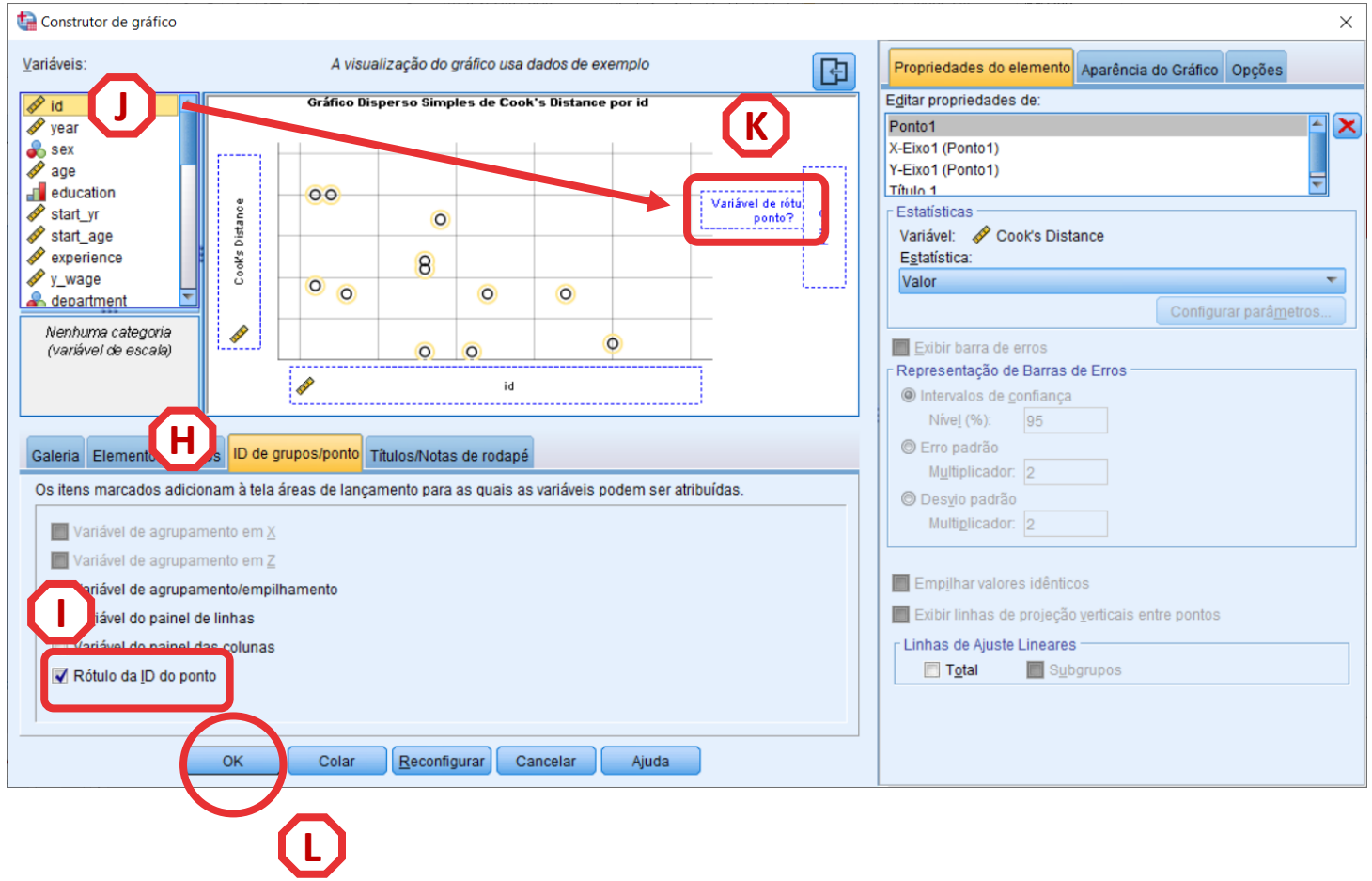
Ausência de Observações Influentes

- Selecionar ‘Gráficos’ / ‘Construtor de Gráfico’ **A**
- Selecionar ‘DispersãoPontos’ **B**
- Selecionar ‘Dispersão (Simples)’ **C**
- Selecionar Variável ‘id’ **D**
- Colocar no eixo ‘x’ **E**
- Selecionar Variável ‘Cooks Distance’ **F**
- Colocar no eixo ‘x’ **G**



Ausência de Observações Influentes

- Selecionar ‘ID de grupos/ponto’ H
- Selecionar ‘Rótulo da ID do Ponto’ I
- Selecionar Variável ‘id’ J
- Colocar na caixa ‘Variável do rótulo do ponto’ K
- Selecionar ‘OK’ L



Ausência de Observações Influentes

- **Reparem que o gráfico permite identificar o ID dos outliers**

- **Interpretação**

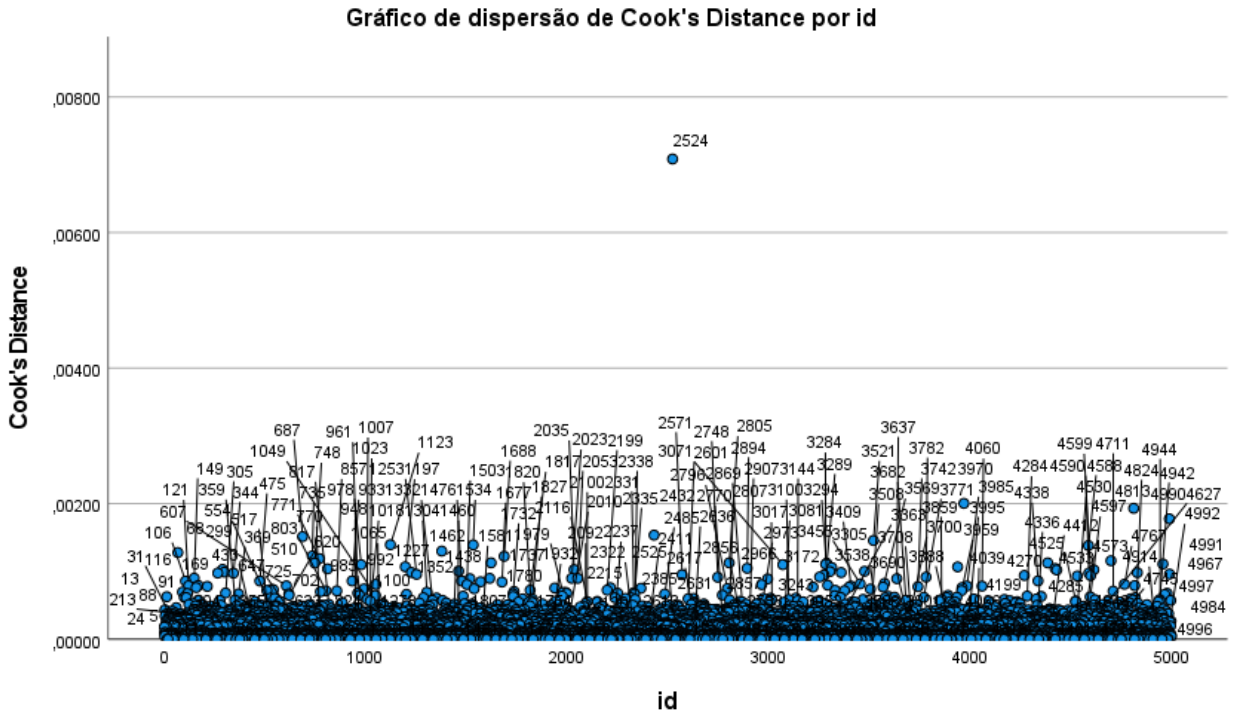
CD > 4/n -> Caso Influente

- **Neste caso, o valor de corte é**

  0.008 (4 / 5000)

- **Neste caso não há observações acima do valor de corte.**

- **Cumpre-se o pressuposto da ausência de observações influentes**



Reportar os resultados do estudo dos pressupostos

Foram realizados uma série de análises para averiguar a adequabilidade do modelo de regressão linear para o estudo destas relações, sendo que todos os pressupostos assumidos com a aplicação deste técnica foram validados. Em primeiro lugar, analisou-se graficamente a linearidade das relações entre as variáveis independentes (experiência e desempenho) com a variável dependente, tendo sido possível observar relações tendencialmente lineares, especialmente entre desempenho e rendimento (Figura x). Apurou-se também a ausência de multicolinearidade entre as variáveis independentes e de controlo com recurso às medidas VIF (<3). Posteriormente, analisou-se a distribuição dos resíduos do modelo, observando-se uma distribuição normal, com um média em torno do valor zero, e com uma variância relativamente constante ao longo dos valores previstos do modelo (Figura x). Os resultados do teste Durbin-Watson, sugerem a ausência de autocorreção significativa nos resíduos (D-W=2) . Apurou-se ainda a existência de observações influentes com a Distância de Cook, admitindo os valores acima de 0,008 (4/N) como indicadores de observações influentes, não tendo sido detetados casos potencialmente problemáticos à estimação do modelo.

Ainda com tempo?

Repetimos o exercício com o caso da aula passada,

incluído a educação no modelo.