Análise de Regressão Linear Múltipla

Análise de Regressão Linear Múltipla

O modelo de regressão linear múltipla postula a existência de uma relação linear entre uma variável dependente ou explicada (Y) e p variáveis independentes ou explicativas $X_1, \ldots, X_p$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

- $X_1, \ldots, X_p$ as variáveis explicativas ou independentes medidas sem erro (não aleatórias);
- ε a variável aleatória residual na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento das variáveis $X_1, \ldots, X_p$ e os possíveis erros de medição;
- $\beta_0, \beta_1, \ldots, \beta_p$ os paraâmetros desconhecidos do modelo (a estimar);

Análise de Regressão Linear Múltipla

Temos então n variáveis aleatórias

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\
Y_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \ldots + \beta_p x_{3p} + \varepsilon_3 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \ldots + \beta_p x_{np} + \varepsilon_n
\end{aligned}$$

Em notação matricial temos

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\ 1 & x_{31} & x_{32} & x_{33} & \ldots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \ldots & x_{np} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Análise de Regressão Linear Múltipla

Estimação dos Parâmetros

A metodologia que se utiliza na regressão linear múltiple é similar que a estudada na regressão linear simples.

A melhor equação ajustada desta forma será aquela que faça mínima a soma de quadrados das desviações das Y observadas e das Y estimadas, onde o modelo estimado é

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \ldots + \hat{\beta}_p X_p$$

Análise de Regressão Linear Múltipla

Assim o sistema de equações normais é

$$(\mathbf{X}'\mathbf{X})\hat{\beta} \quad = \quad \mathbf{X}'\mathbf{Y}$$

da onde resolvendo para $\hat{\beta}$ obtemos

$$\hat{\beta} \quad = \quad (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

aqui $(\mathbf{X}'\mathbf{X})^{-1}$ representa a matriz inversa de $(\mathbf{X}'\mathbf{X})$. Notar que $(\mathbf{X}'\mathbf{X})$ é simêtrica.

Análise de Regressão Linear Múltipla

Propriedades do Estimador $\hat{\beta}$

Em forma similar ao caso simples, o estimador de mínimos quadrados tem as seguintes propriedades:

1. $\hat{\beta}$ é não viciado, ou seja $E(\hat{\beta}) = \beta$, isto significa que para todo $j = 0, 1, \cdots, p$, $E(\hat{\beta}_j) = \beta_j$
2. $\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
3. O estimador mínimo-quadrático $\hat{\beta}$ é o melhor estimador linear de β no sentido que é o que tem a menor variância.

Análise de Regressão Linear Múltipla

Estimação da Variância σ^2

No modelo de regressão linear múltiple com p variáveis preditoras (com o intercepto há no total p+1 parâmetros a estimar), o estimador da variância dos errors esta dado por:

$$\hat{\sigma}^2 = \frac{\sum \varepsilon_i^2}{n-p-1} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\beta})'(\mathbf{Y}-\mathbf{X}\hat{\beta})}{n-p-1} = \frac{SQE}{n-p-1}$$

O numerador da expressão representa a soma de quadrados dos resíduos (uma das fontes de variação da análise de variância)

Análise de Regressão Linear Múltipla

Testes sobre os coeficientes de regressão

Queremos testar se a influência de uma variável explicativa sobre a variável dependente não é significativa. Para isso testamos a hipótese nula que o coeficiente para esta variável é nulo:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

Se H_0 for rejeitada então temos evidência de que $\beta_i \neq 0$, isto é a variável explicativa X_i é útil na predição do valor da variável dependente.
Se H_0 não for rejeitada então a variável explicativa X_i é geralmente retirada da equação de regressão pois não influência significativamente a variável resposta Y.

Análise de Regressão Linear Múltipla

Os resultados podem ser convenientemente resumidos na tabela da ANOVA.

ANOVA

Fontes de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	Razão F
devido à Regressão	SQR	p	QMR	
devido aos resíduos	SQE	n-p-1	QME	$F = \frac{QMR}{QME}$
total	SQT	n-1		

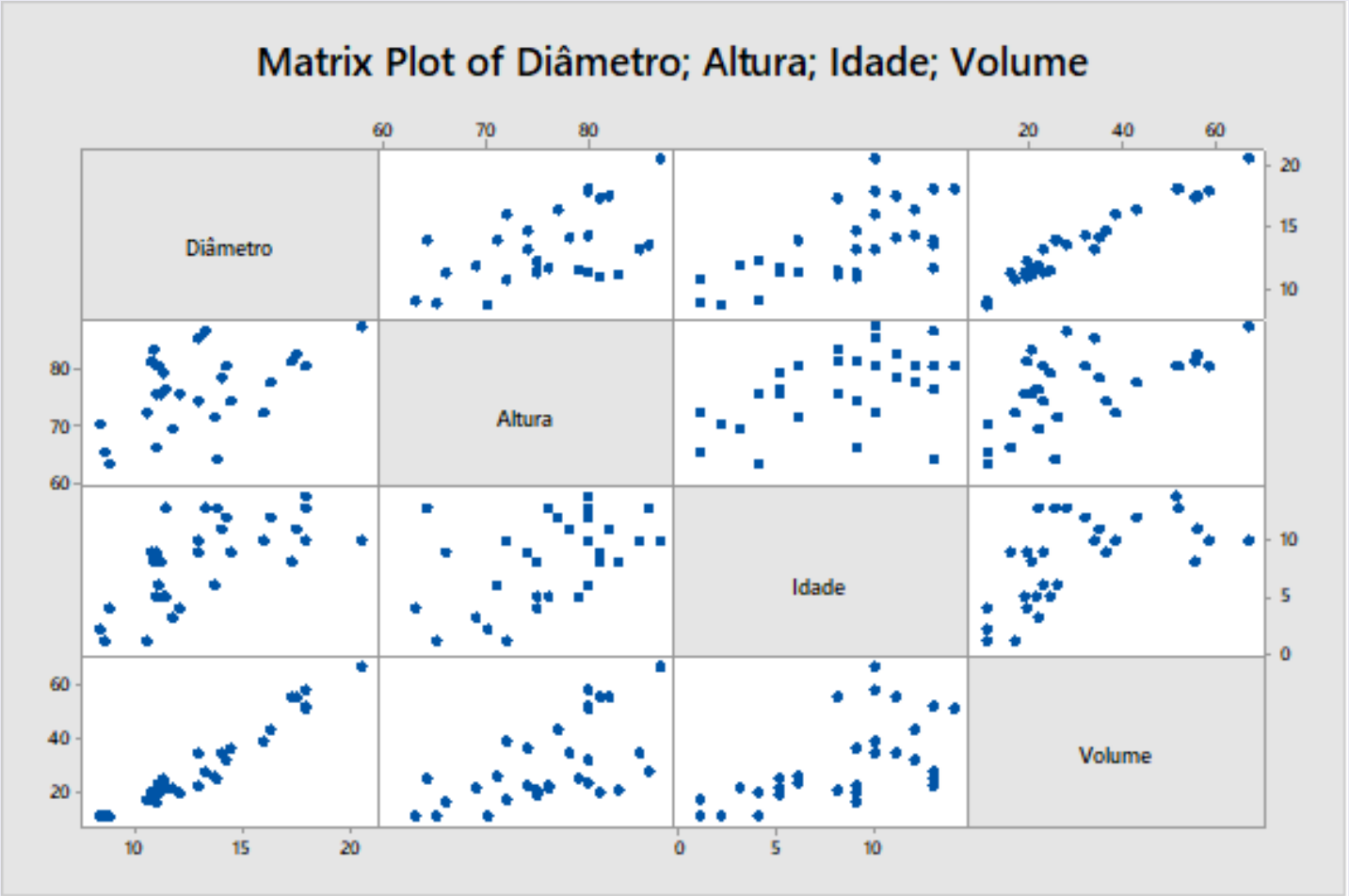
Exemplo 1

Se deseja construir um modelo de regressão para obter o volume de madeira de uma cerejeira em função da altura do tronco e do diâmetro do mesmo a um metro do solo e da idade da árvore. Foi tomada uma amostra de 31 árvores.

Obs	Diâmet	Alt	Ida	Vol	Obs	Diâmet	Alt	Ida	Vol
1	8,3	70	2	10,3	17	12,9	85	10	33,8
2	8,6	65	1	10,3	18	13,3	86	13	27,4
3	8,8	63	4	10,2	19	13,7	71	6	25,7
4	10,5	72	1	16,4	20	13,8	64	13	24,9
5	10,7	81	9	18,8	21	14	78	11	34,5
6	10,8	83	8	19,7	22	14,2	80	12	31,7
7	11	66	9	15,6	23	14,5	74	9	36,3
8	11	75	5	18,2	24	16	72	10	38,3
9	11,1	80	6	22,6	25	16,3	77	12	42,6
10	11,2	75	8	19,9	26	17,3	81	8	55,4
11	11,3	79	5	24,2	27	17,5	82	11	55,7
12	11,4	76	5	21	28	17,9	80	10	58,3
13	11,4	76	13	21,4	29	18	80	13	51,5
14	11,7	69	3	21,3	30	18	80	14	51
15	12	75	4	19,1	31	20,6	87	10	67
16	12,9	74	9	22,2					

Exemplo 1

Diagrama de Dispersão



Exemplo 1

Ajuste do Modelo

Regression Analysis: Volume versus Diâmetro; Altura; Idade

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	6990,4	2330,13	228,04	0,000
Error	27	275,9	10,22		
Total	30	7266,3			

Model Summary

S	R-sq	R-sq(adj)
3,19656	96,20%	95,78%

Coefficients

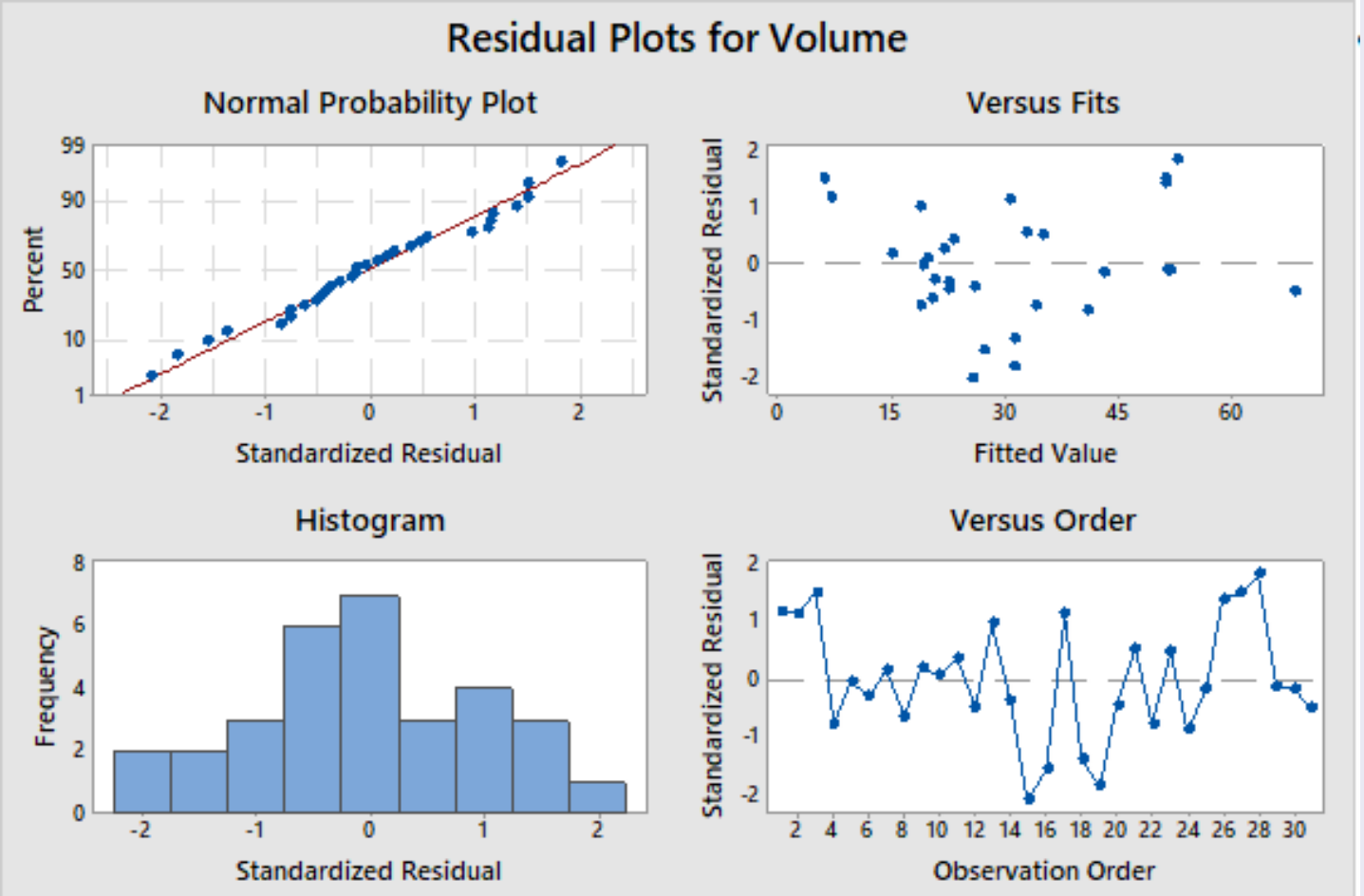
Term	Coef	SE Coef	T-Value	P-Value
Constant	-56,84	7,33	-7,75	0,000
Diâmetro	4,829	0,261	18,48	0,000
Altura	0,350	0,109	3,21	0,003
Idade	-0,478	0,209	-2,29	0,030

Regression Equation

Volume = -56,84 + 4,829 Diâmetro + 0,350 Altura - 0,478 Idade

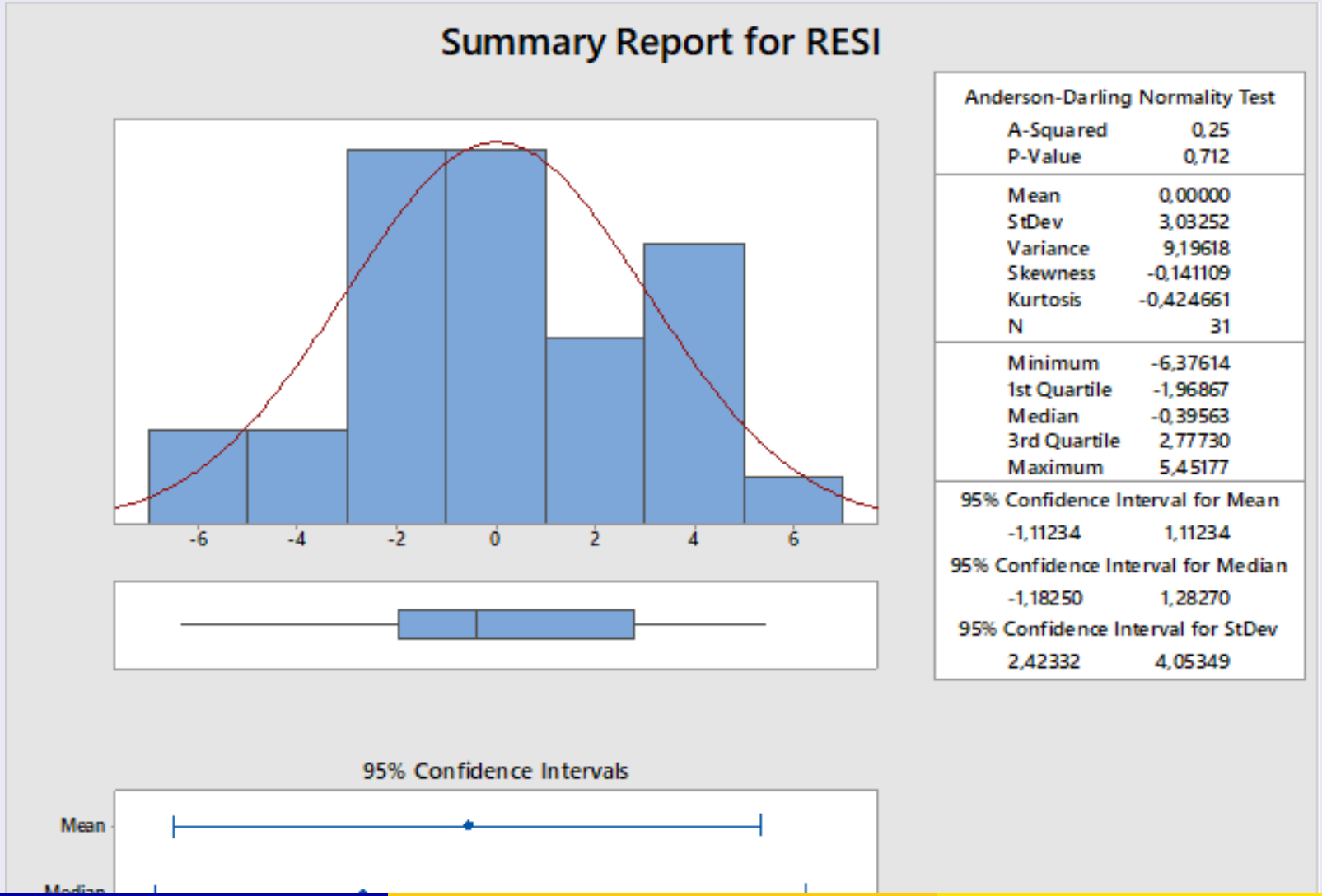
Exemplo 1

Análise de Resíduos



Exemplo 1

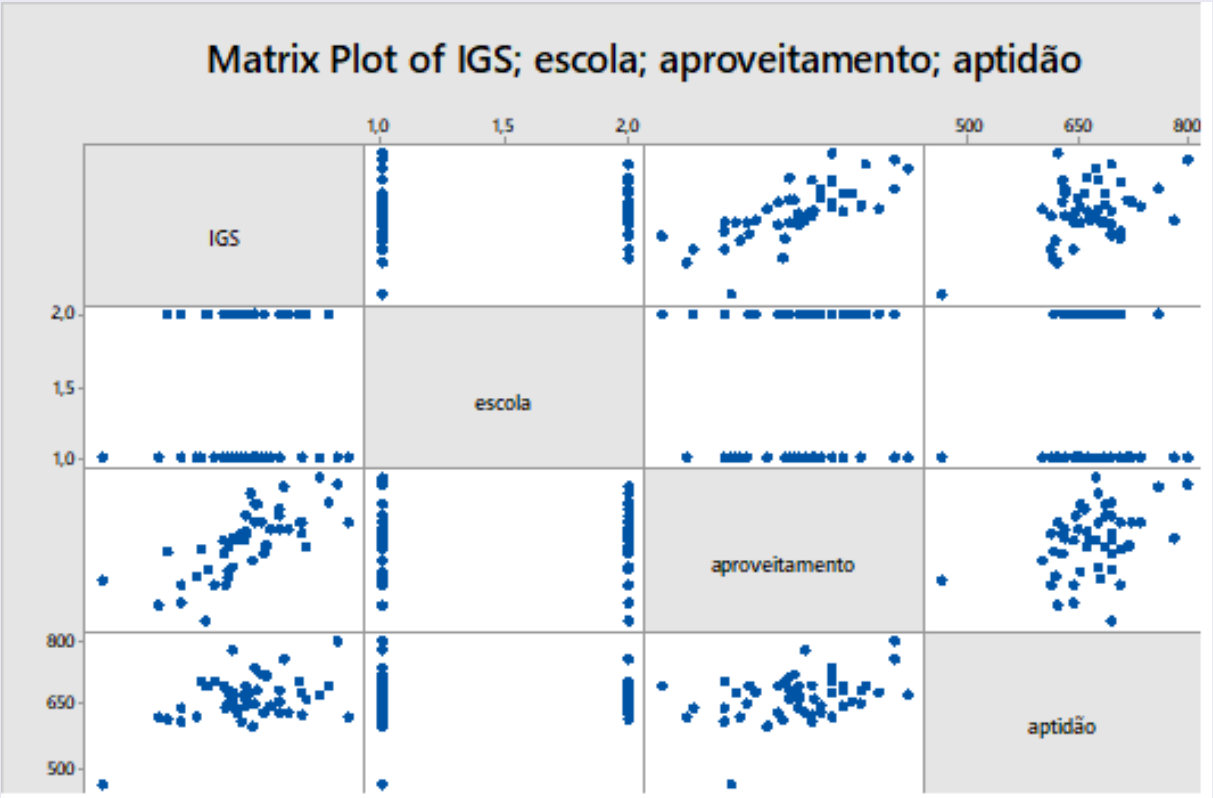
Normalidade dos Erros



Exemplo 2

Num estudo feito em 50 estudantes, na Universidade de Porto Rico, mostrou-se que o Indice Geral dos Estudantes estão em função da pontuação de aptidão à matemática, aproveitamento matemático e tipo de escola (1 pública e 2 privada). Escolher um modelo e justificar.

Diagrama de Dispersão



Exemplo 2

Ajuste do Modelo 1

Regression Analysis: IGS versus escola; aproveitamento; aptidão

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	6952	2317,3	19,54	0,000
Error	46	5455	118,6		
Total	49	12407			

Model Summary

S	R-sq	R-sq(adj)
10,8896	56,03%	53,17%

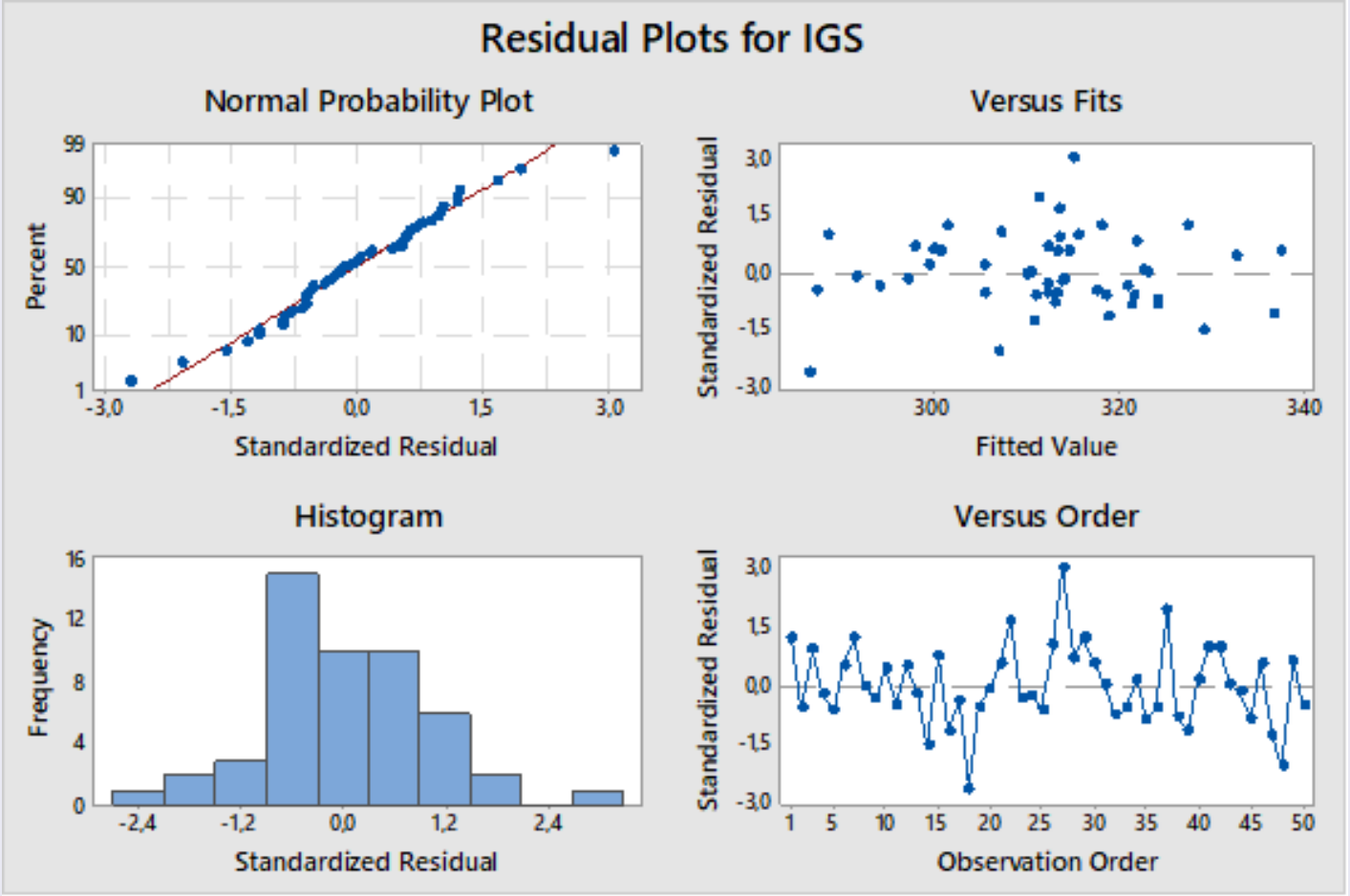
Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	135,9	24,5	5,55	0,000
escola	1,93	3,09	0,63	0,535
aproveitamento	0,1970	0,0315	6,25	0,000
aptidão	0,0569	0,0314	1,81	0,077

Regression Equation

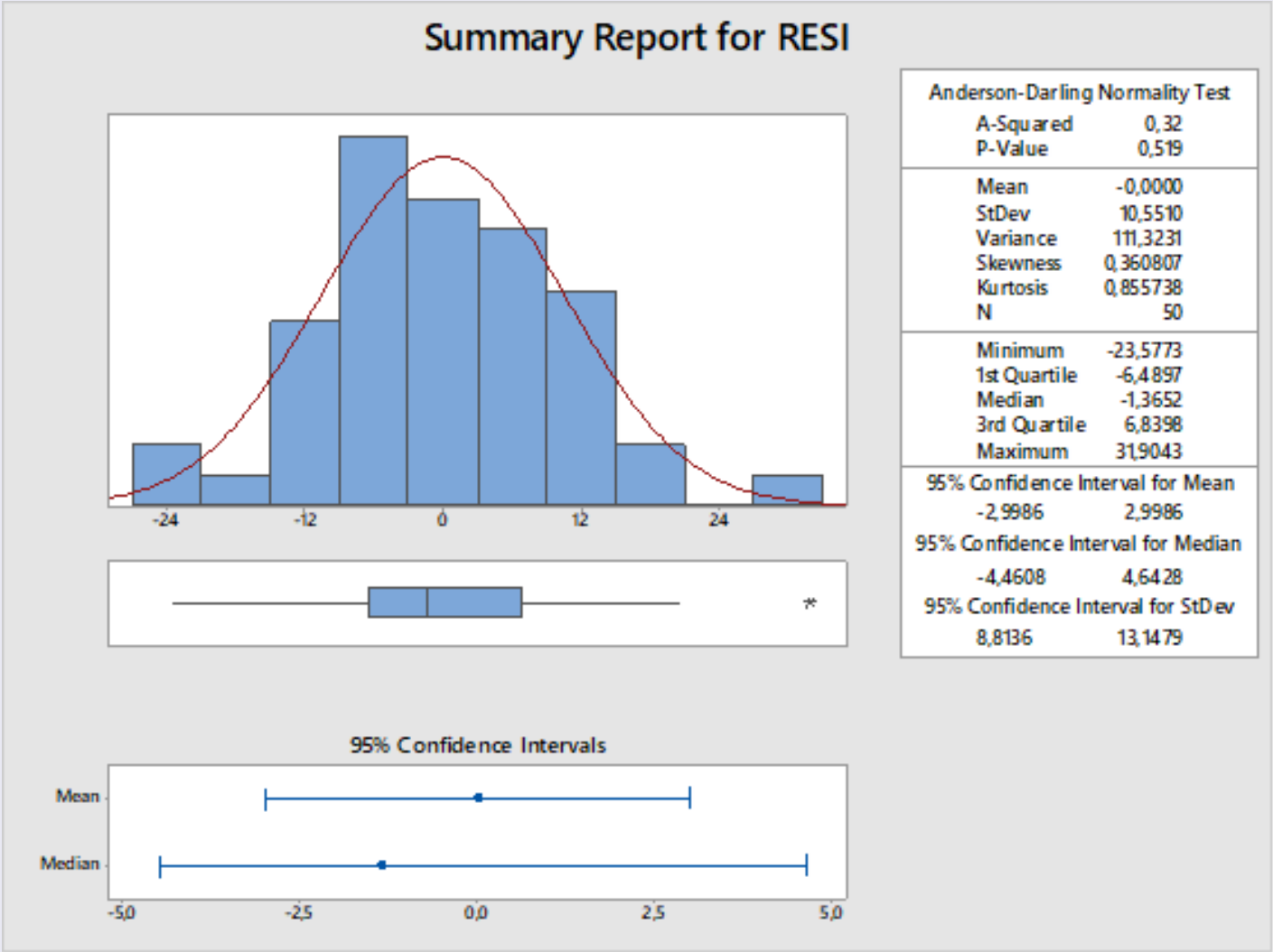
Exemplo 2

Análise de Resíduos



Exemplo 2

Normalidade dos Erros



Anderson-Darling Normality Test	
A-Squared	0,32
P-Value	0,519
Mean	-0,0000
StDev	10,5510
Variance	111,3231
Skewness	0,360807
Kurtosis	0,855738
N	50
Minimum	-23,5773
1st Quartile	-6,4897
Median	-1,3652
3rd Quartile	6,8398
Maximum	31,9043
95% Confidence Interval for Mean	
-2,9986	2,9986
95% Confidence Interval for Median	
-4,4608	4,6428
95% Confidence Interval for StDev	
8,8136	13,1479

Exemplo 2

Ajuste do Modelo 2

Regression Analysis: IGS versus aproveitamento; aptidão

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	6906	3452,8	29,50	0,000
Error	47	5501	117,0		
Total	49	12407			

Model Summary

S	R-sq	R-sq(adj)
10,8188	55,66%	53,77%

Coefficients

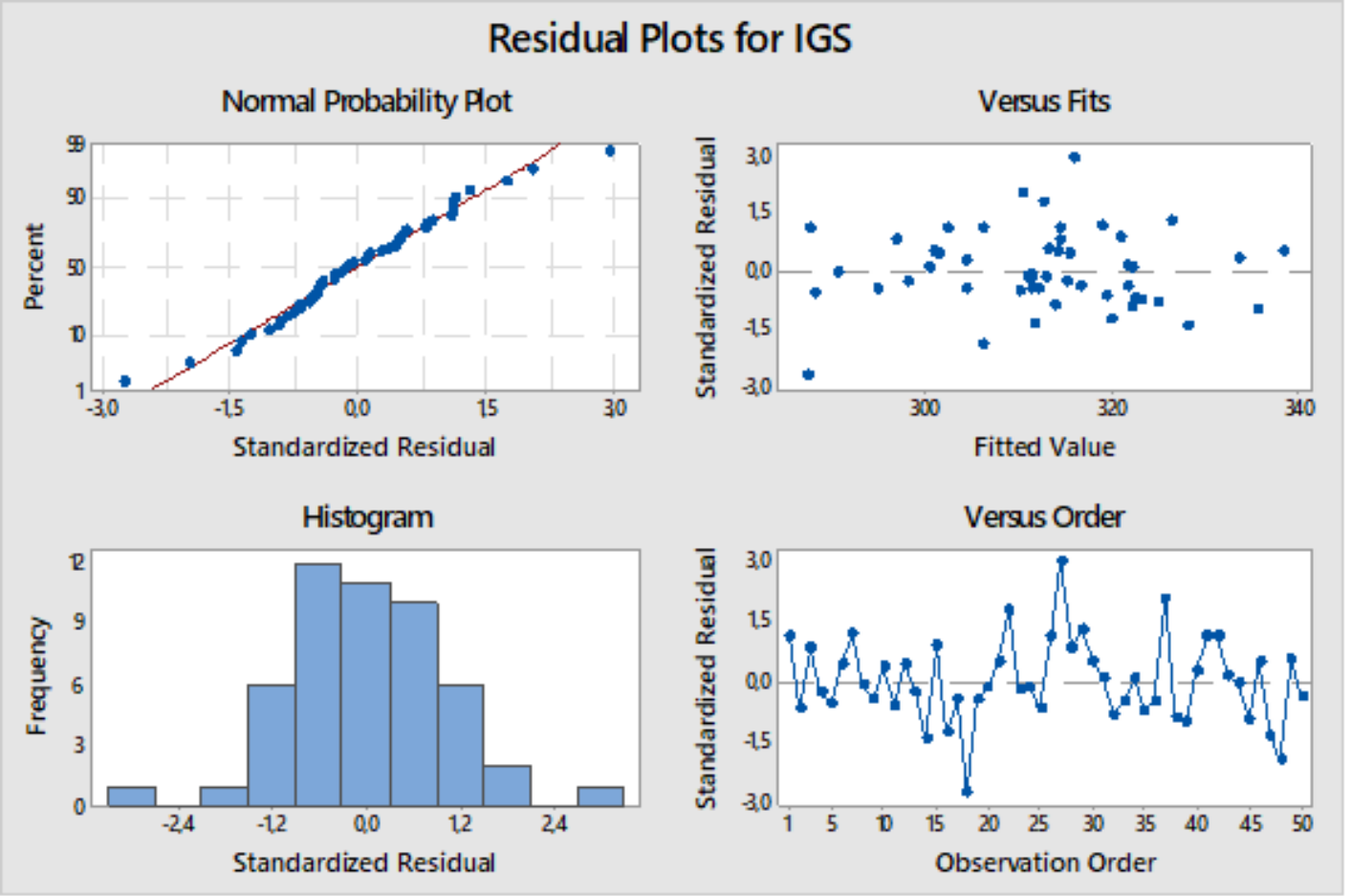
Term	Coef	SE Coef	T-Value	P-Value
Constant	138,6	24,0	5,78	0,000
aproveitamento	0,1969	0,0313	6,29	0,000
aptidão	0,0573	0,0312	1,84	0,073

Regression Equation

IGS = 138,6 + 0,1969 aproveitamento + 0,0573 aptidão

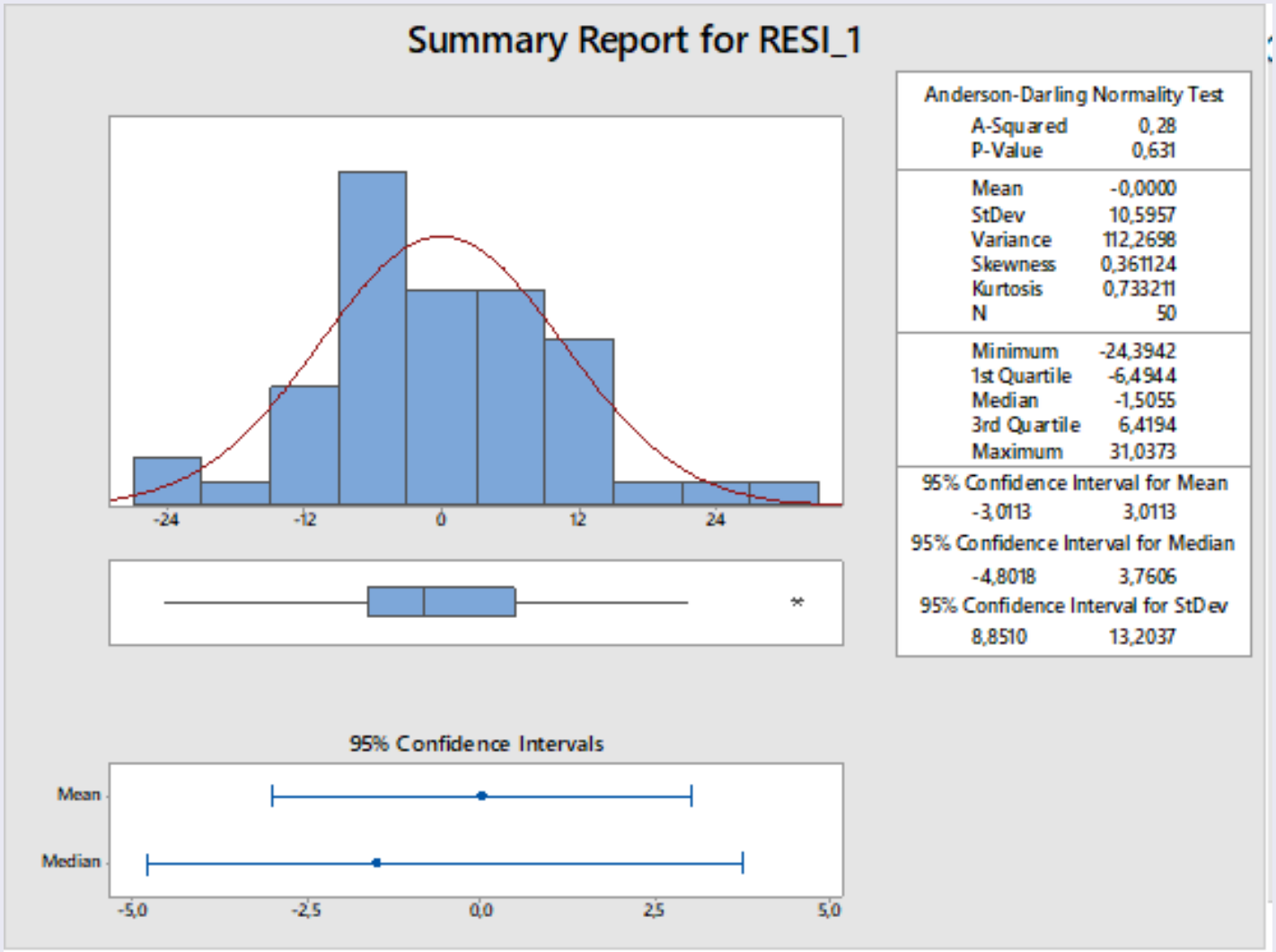
Exemplo 2

Análise de Resíduos



Exemplo 2

Normalidade dos Erros



Exemplo 2

Ajuste do Modelo 3

Regression Analysis: IGS versus aproveitamento

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	6511	6511,3	53,01	0,000
Error	48	5896	122,8		
Total	49	12407			

Model Summary

S	R-sq	R-sq(adj)
11,0826	52,48%	51,49%

Coefficients

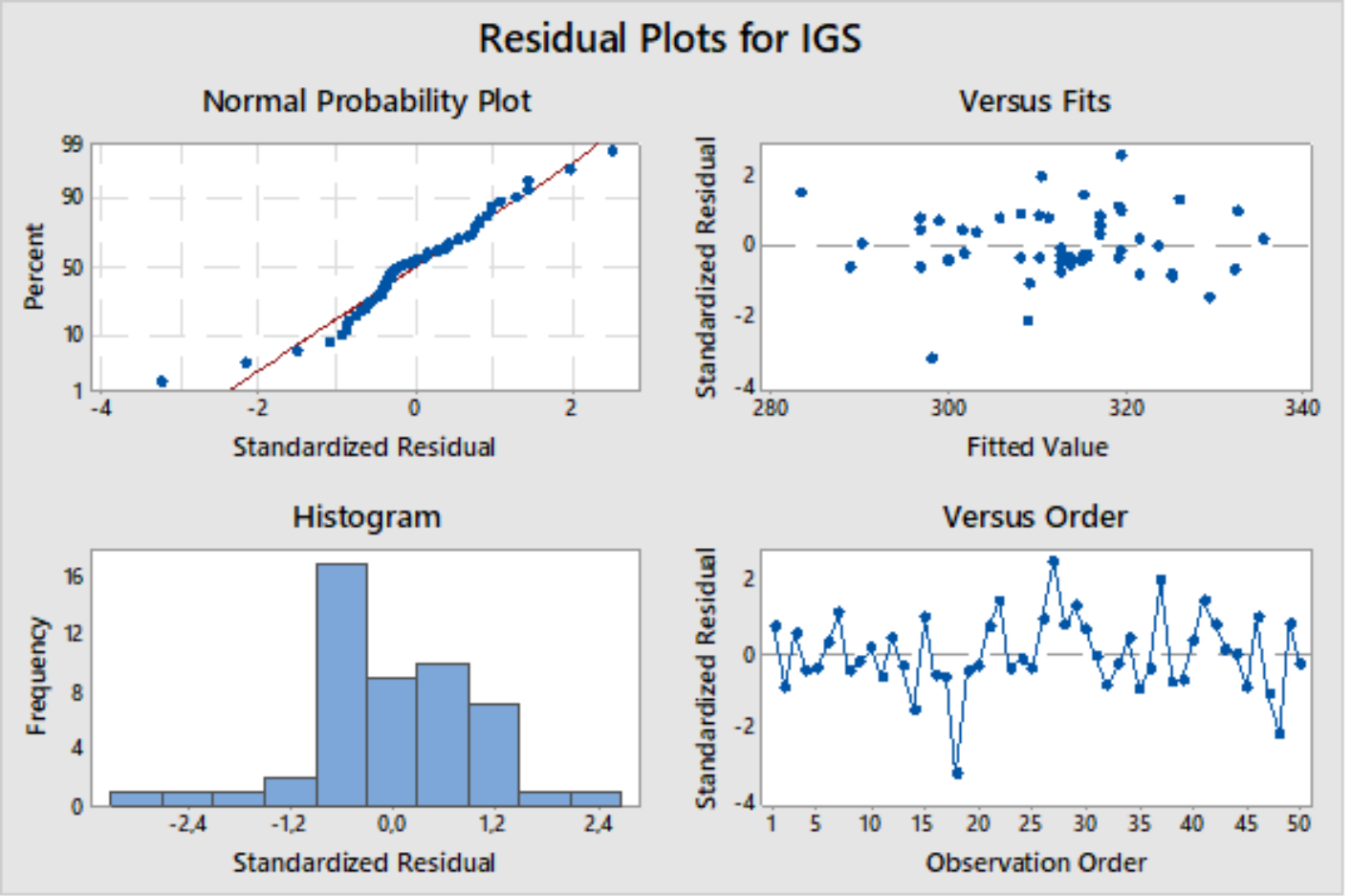
Term	Coef	SE Coef	T-Value	P-Value
Constant	162,4	20,6	7,87	0,000
aproveitamento	0,2177	0,0299	7,28	0,000

Regression Equation

IGS = 162,4 + 0,2177 aproveitamento

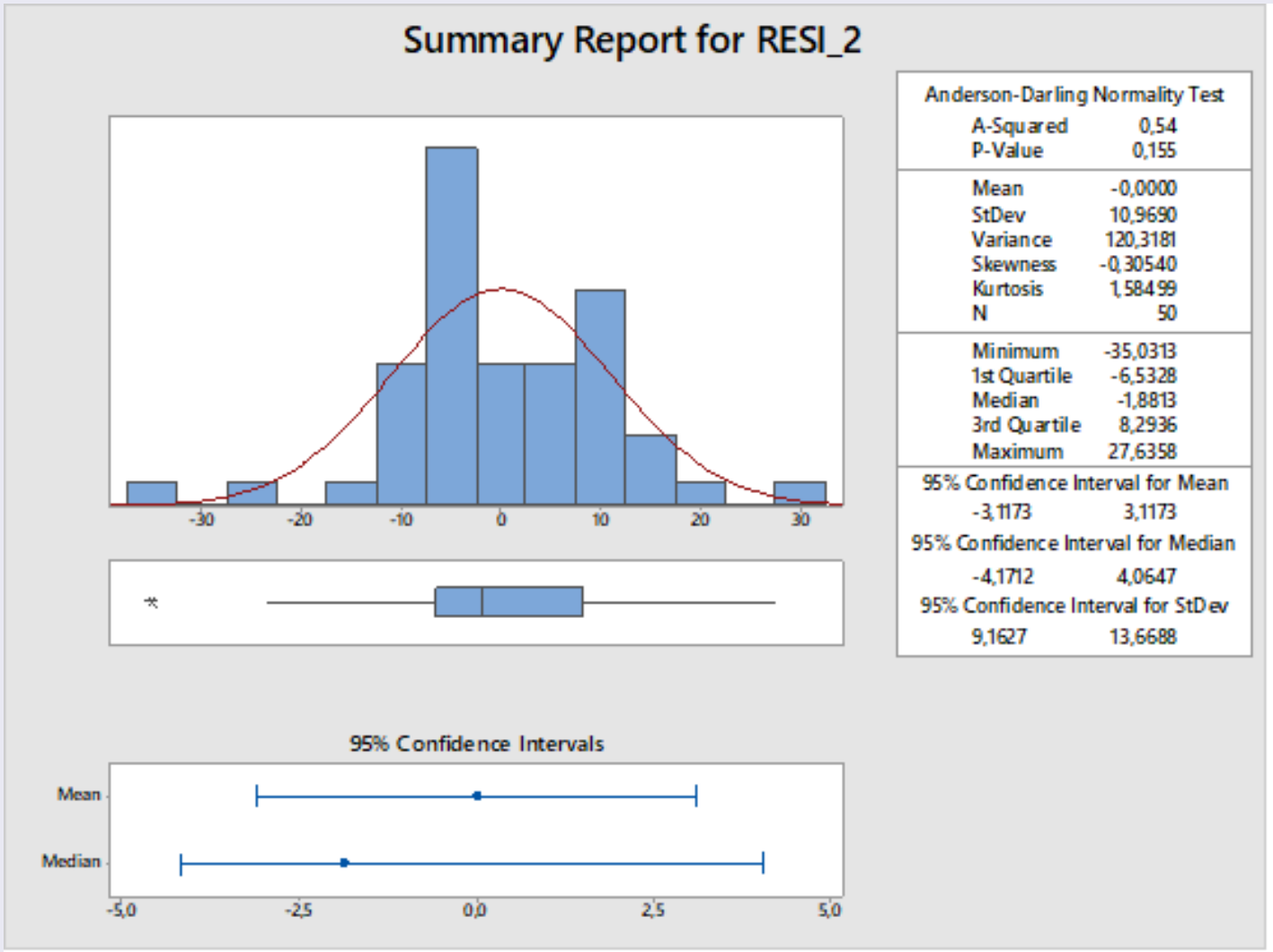
Exemplo 2

Análise de Resíduos



Exemplo 2

Normalidade dos Erros



Anderson-Darling Normality Test	
A-Squared	0,54
P-Value	0,155
Mean	-0,0000
StDev	10,9690
Variance	120,3181
Skewness	-0,30540
Kurtosis	1,58499
N	50
Minimum	-35,0313
1st Quartile	-6,5328
Median	-1,8813
3rd Quartile	8,2936
Maximum	27,6358
95% Confidence Interval for Mean	
-3,1173	3,1173
95% Confidence Interval for Median	
-4,1712	4,0647
95% Confidence Interval for StDev	
9,1627	13,6688