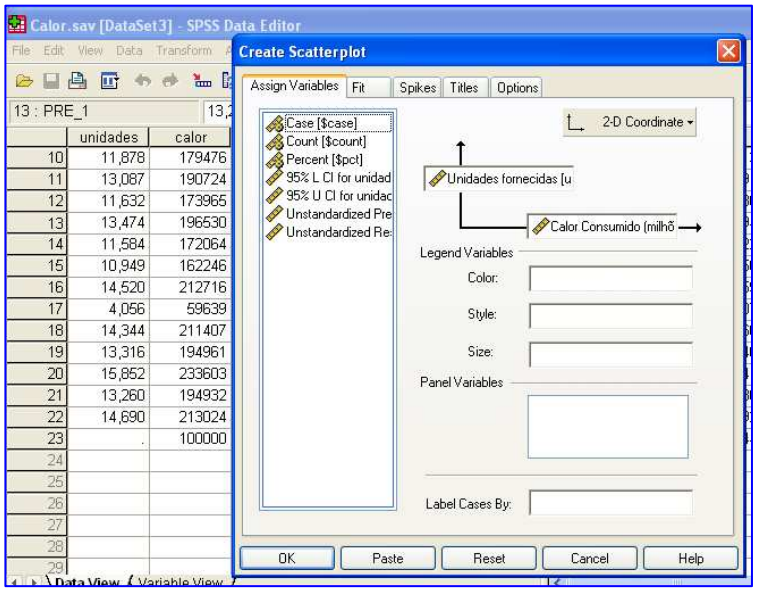
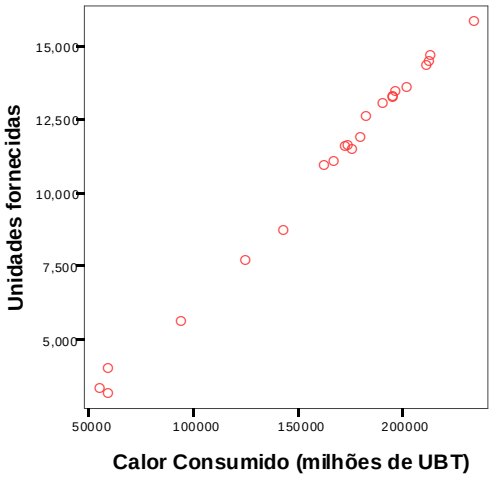
Regressão Linear em SPSS

1. No ficheiro **Calor.sav** encontram-se os valores do consumo mensal de energia, medido em milhões de unidades termais britânicas, acompanhados de valores de output, em milhões de kWh, de electricidade fornecida por uma central termo-eléctrica em Inglaterra.

   1.1. Construa um gráfico de dispersão que permita relacionar ambas as variáveis com a intenção de identificar uma possível relação linear.

em SPSS: Graph / Interactive /Scatterplot



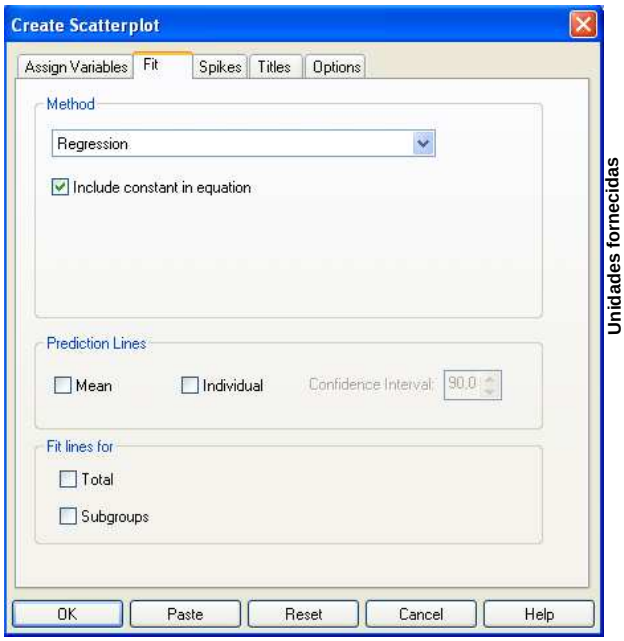


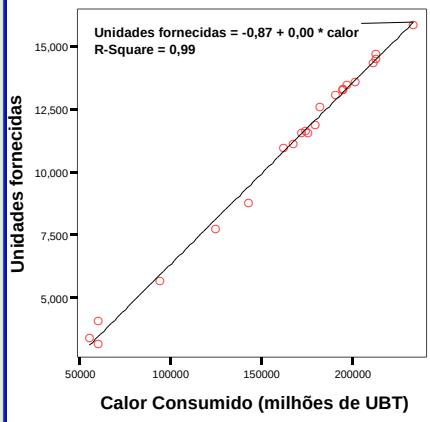
Da observação do gráfico de dispersão é razoável afirmar que existe uma relação linear entre as duas variáveis.

1.2. Estabeleça o modelo a ajustar aos dados

Como do gráfico de dispersão podemos constatar que existe uma relação linear entre as duas variáveis podemos usar um **modelo de regressão linear** para ajustar estes dados.

Note que se seleccionamos no menu: **Graph / Interactive /Scatterplot, o** tab **Fit** como método para ajustar os dados **Regression** podemos obter o gráfico de dispersão com a recta de regressão desenhada e a sua equação.



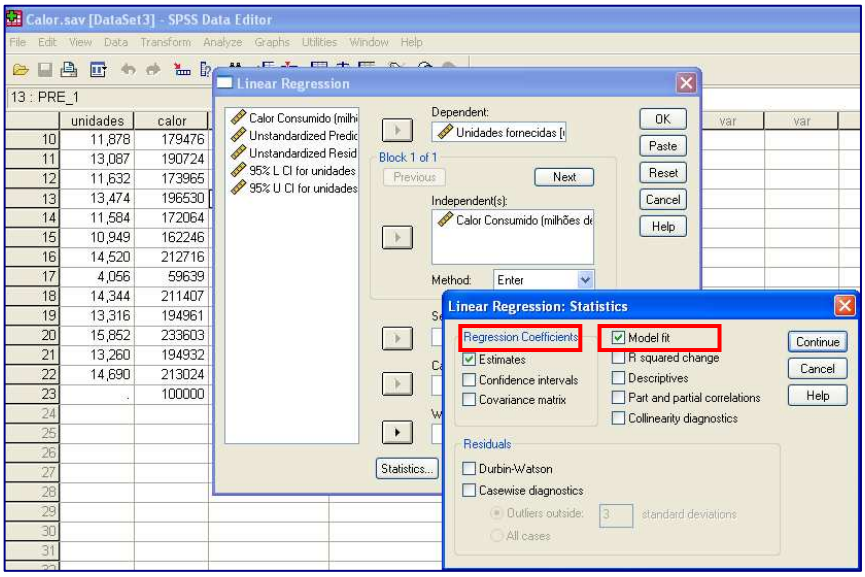


Linear Regression

Note que o valor do declive na recta de regressão é 0.00, mas isto é devido à aproximação usada. Como poderemos verificar logo este valor é diferente de 0, porem é um valor muito pequeno, da ordem de 10^{-5}

Mas de uma forma mais geral, a análise de regressão linear no SPSS é efectuada através do menu:

em SPSS: Analize / Regression /Linear



O método do mínimo dos quadrados é o método implementado em SPSS para estimar os coeficientes de regressão Com as opções do SPSS seleccionadas podemos obter como output a seguintes 4 tabelas:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Calor Consumido (milhões de UBT)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Unidades fornecidas

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,997(a)	**,995**	,994	,282649

a Predictors: (Constant), Calor Consumido (milhões de UBT)

O coeficiente de correlação **R=0, 997 ≈ 1**, pelo que é evidente a existência de uma relação linear entre as variáveis em estudo

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	302,731	1	302,731	3789,321	,000(a)
	Residual	1,598	20	,080		
	Total	304,329	21			

a Predictors: (Constant), Calor Consumido (milhões de UBT)

b Dependent Variable: Unidades fornecidas

O teste realizado pela ANOVA é:
H_0: $b_1 = 0$ vs. H_1: $b_1 \neq 0$
Como o **p-value=0** $\Rightarrow$ para q.q. nível de significância rejeita-se H_0
$\Rightarrow b_1 \neq 0 \Rightarrow$ a regressão linear tem significado para q.q. nível de significância

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	**-,869**	,201		-4,329	**,000**
	Calor Consumido (milhões de UBT)	**7,20E-005**	,000	,997	61,557	**,000**

a Dependent Variable: Unidades fornecidas

Valores observados das estatísticas dos testes:
para a ordenada na origem b_0:
t_{0obs} = -4, 328 $(T_0 \sim t_{n-2})$
para o declive b_1:
t_{0obs} = 61, 328 $(T_1 \sim t_{n-2})$

p-value para a ordenada na origem: 0
p-value para o declive: 0

Modelo de Regressão Linear:
unid. fornecidas = $b_0 + b_1$ x calor consumido + **ε**
com erro **ε** ~ N (0, σ2)
Estimativas dos coeficientes:

b_0 = -0, 869, $b_1 = 7{,}20 \times 10^{-5}$

1.3. Com base nos resultados obtidos responda as seguintes questões:

a. Quais as estimativas do declive (b_1) e da ordenada na origem (b_0) da recta de regressão?

$\hat{b}_0 =$ -0, 869 $\qquad$ $\hat{b}_1 = 7{,}20 \times 10^{-5}$

b. Qual a equação da recta de regressão?

$y = -0{,}869 + 7{,}20 \times 10^{-5}\ x$

c. O valor do declive é significativamente diferente de 0, ao nível de significância 5%?

i. Escreva as hipóteses em causa

$H_0: b_1 = 0$ vs $H_1: b_1 \neq 0$

ii. Indique o valor do p-value do teste
p-value = 0

iii. Conclua
A hipótese nula é rejeitada para q.q nível de significância. Conclui-se que o declive não é nulo para q.q. nível de significância

d. A ordenada na origem é significativamente diferente de 0, ao nível de significância 5%?

i. Escreva as hipóteses em causa

$H_0: b_0 = 0$ vs $H_1: b_0 \neq 0$

ii. Indique o valor do p-value do teste: 0

iii. Conclua: A hipótese nula é rejeitada para q.q nível de significância. Conclui-se que a ordenada na origem não é nula para q.q. nível de significância

1.4. Efectue os cálculos necessários para obter os p-values dos testes para os coeficientes de regressão mostrados na tabela dos coeficientes

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Unidades fornecidas	22	3,173	15,852	10,91527	3,806819
Calor Consumido (milhões de UBT)	22	55266	233603	163559,41	52698,343
Valid N (listwise)	22				

Da tabela das estatísticas descritivas obtemos n=22

O p-value para um teste bilateral é igual a:

$2P(T < t_{obs} \mid H_0)$ se t_{obs} for reduzido

$2P(T > t_{obs} \mid H0)$ se t_{obs} for elevado

O valor observado da estatística do teste t_{obs} considera-se reduzido (elevado) se a estimativa que se obtém para o parâmetro a testar é inferior (superior) ao valor especificado em H_0

- **Teste de hipótese para a ordenada na origem b_0 da recta de regressão:**
  - H_0: $b_0 = 0$ vs. H_1: $b_0 \neq 0$
  - $t_{0obs} = -4{,}329$ (valor observado da estatística do teste, ver tabela dos coeficientes)
  - o valor observado da estatística do teste é reduzido pois a estimativa que se obtém para b_0 (-0.869) é um valor inferior a 0 (o valor especificado em H_0).

  Assim:
  p-value = 2 P(T<-4.329) = 2 t_{n-2}(-4.329) = 2 (1- t_{n-2}(4.329))
  = 2 (1- CDF.T(4.329, 20)) = 2 x 0 = 0

- **Teste de hipótese para o declive b_1 da recta de regressão:**
  - H_0: $b_1 = 0$ vs. H_1: $b_1 \neq 0$
  - $t_{1obs} = 61{,}777$ (valor observado da estatística do teste, ver tabela dos coeficientes)
  - o valor observado da estatística do teste é elevado pois a estimativa que se obtém para b_1 ($7{,}20 \times 10^{-5}$) é um valor superior a 0 (o valor especificado em H_0).

  Assim:
  p-value = 2 P(T>61.777) = 2 (1- P(T <61.557) = 2 (1- t_{n-2}(61.557))
  = 2 (1-CDF.T(61.557, 20)) = 2 x 0 = 0

1.5. Qual é a proporção de variabilidade de Y explicada por x?

Da tabela de ANOVA podemos obter o coeficiente de determinação $R^2 = ,995$ (ver R square). Este coeficiente mede a quantidade de variabilidade explicada por x, isto é, pelo modelo de regressão já que consiste na razão entre a soma dos quadrados devido aos resíduos (SS_R) e a soma dos quadrados total (S_{YY}).

$$R^2 = \frac{S_{xY}^2}{S_{xx}S_{YY}} = \frac{SS_R}{S_{YY}} = 1 - \frac{SS_E}{S_{YY}}$$

$$SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Então, $R^2 = ,995$ quer dizer que 99.5% da variabilidade encontrada para y é explicada por x e apenas os restantes 0,5% se devem a outros factores.
Um bom ajuste do modelo deve reflectir-se num valor de R^2 próximo de 1. Como neste caso o coeficiente de determinação é bastante elevado (muito próximo de 1), podemos concluir que a relação linear entre as duas variáveis é forte.

1.6. Proceda à análise dos resíduos com a intenção de validar os pressupostos do modelo.

Pressupostos de regressão: os erros são independentes e identicamente distribuídos com distribuição Normal de media zero e variância σ^2. Uma vez que não conhecemos os erros temos que analisar a sua estimativa que é dada pelos resíduos:

$$e_i = y_i - \hat{b}_0 - \hat{b}_1 x_i = y_i - \hat{y}_i$$

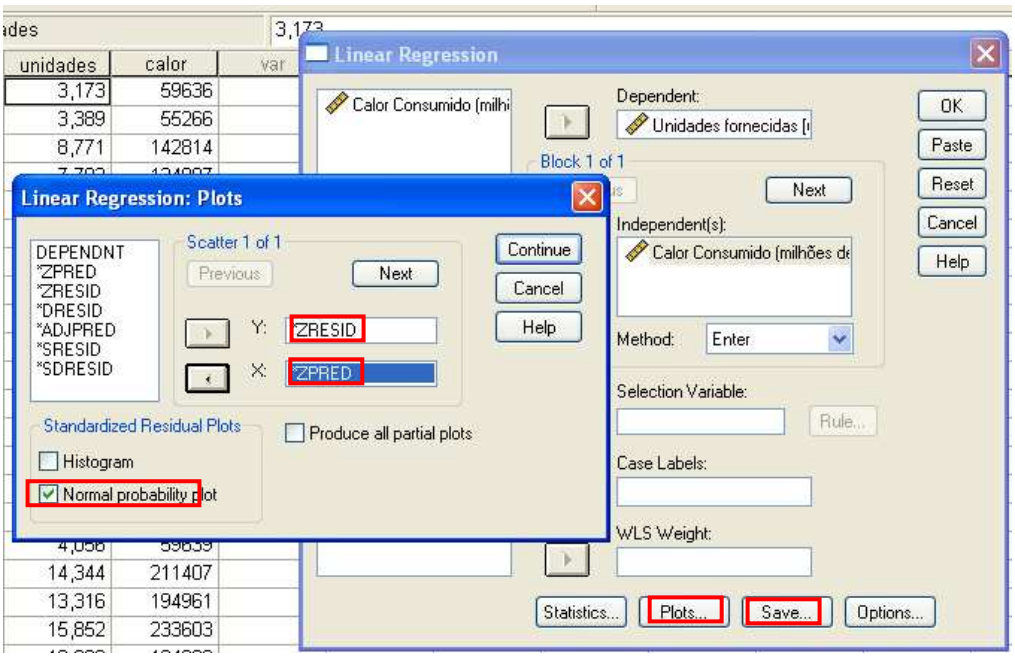
Para validar que os resíduos têm distribuição Normal:

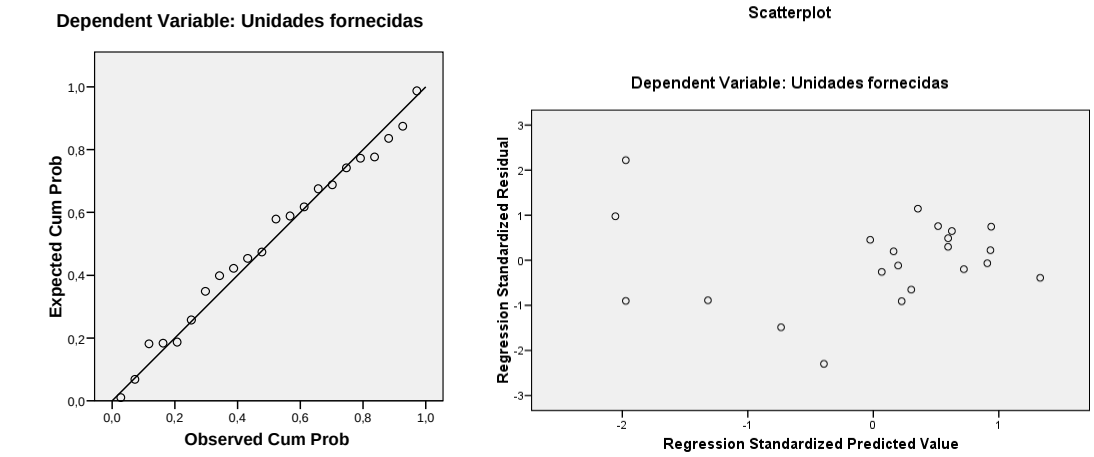
- construir QQ-plot ou PP-plot dos resíduos, realizar teste de ajustamento de Kolmogorov-Smirnov
- através do menu de Regressão Linear podemos fazer directamente um PP-plot dos resíduos

Para validar que os resíduos são independentes e identicamente distribuídos (são aleatórios e com variância constante):

- construir gráficos de resíduos versus valores preditos ou observados.

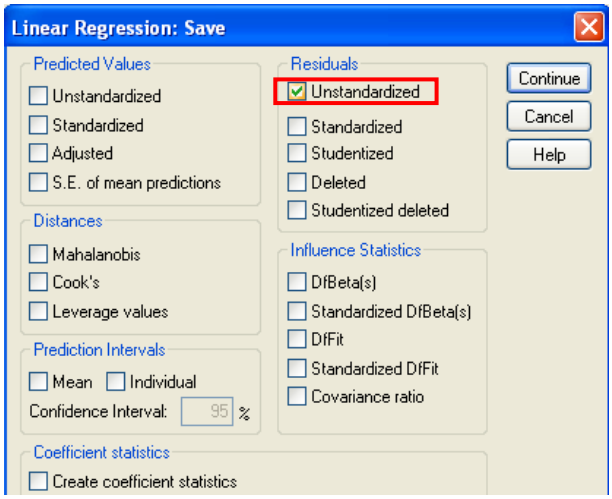
Todos estes gráficos podem ser feitos através do menu de Regressão Linear:



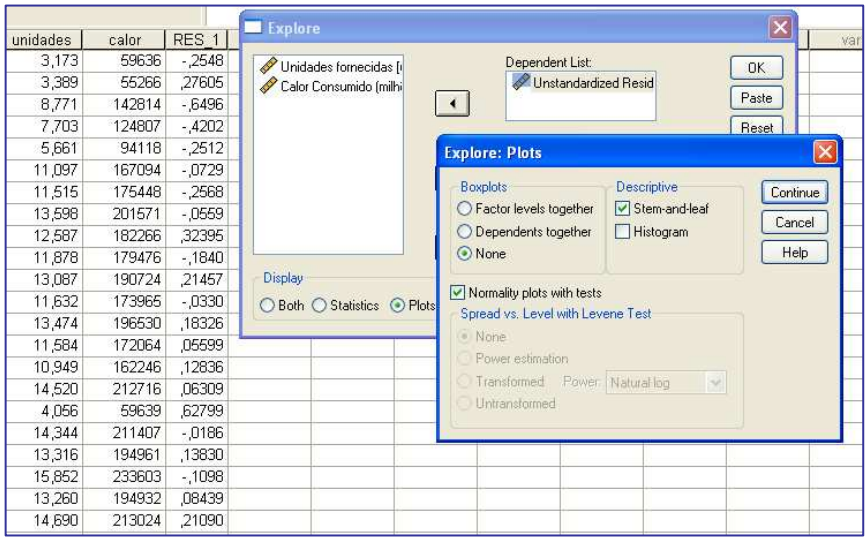


- O PP-plot não nos dá qualquer indicação que contrarie o pressuposto da normalidade dos resíduos
- O gráfico de dispersão dos resíduos em função dos valores preditos estandardizados mostra-se bastante aleatório

Também podemos fazer um QQ-plot ou um teste de ajustamento de K-S para validar os pressupostos de normalidade dos resíduos. Para isto devemos guardar os resíduos numa nova variável, usando a opção **Save** do menu de **Linear Regression**



Depois podemos escolher o menu **Analyze \ Descriptive Statistics \ Explore** com a opção **Normality plots with tests**



Usando a variável RES-1 (os resíduos guardados) e fazendo um QQ-plot e os testes de ajustamento de Kolmogorov-Smirnov e de Shapiro Wilk podemos concluir que os resíduos têm distribuição Normal (o QQ-plot identifica um ajuste entre os quantis amostrais e os quantis de distribuição Normal e os testes de ajustamentos fornecem valores de p-values superiores aos níveis usuais de significância.

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,085	22	**,200(*)**	,982	22	**,940**

* This is a lower bound of the true significance.
a Lilliefors Significance Correction