Esercizio 3
(Regressione con trasformazione di variabili)

DATI

Il data set television.sav (fonte: http://www.statsci.org) contiene dati relativi 40 stati. Le variabili sono

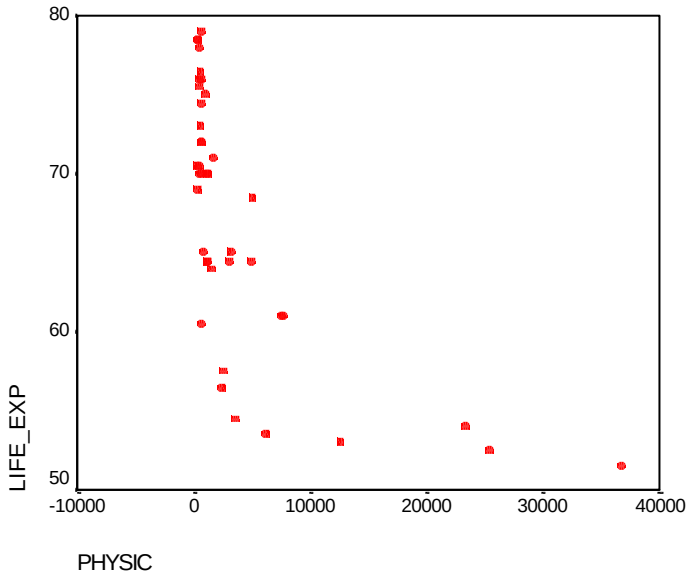
COUNTRY	name of country
LIFE_EXP	Life expectancy at birth (in years)
TEL	People per television
PHYSIC	People per physician
F_LFEXP	Female life expectancy
M_LFEXP	Male life expectancy

Domanda

Ci chiediamo se la speranza di vita alla nascita (LIFE_EX) dipenda dal numero di pazienti per medico (PHYSIC). A tale scopo stimiamo un modello di regressione lineare, tramite SPSS.

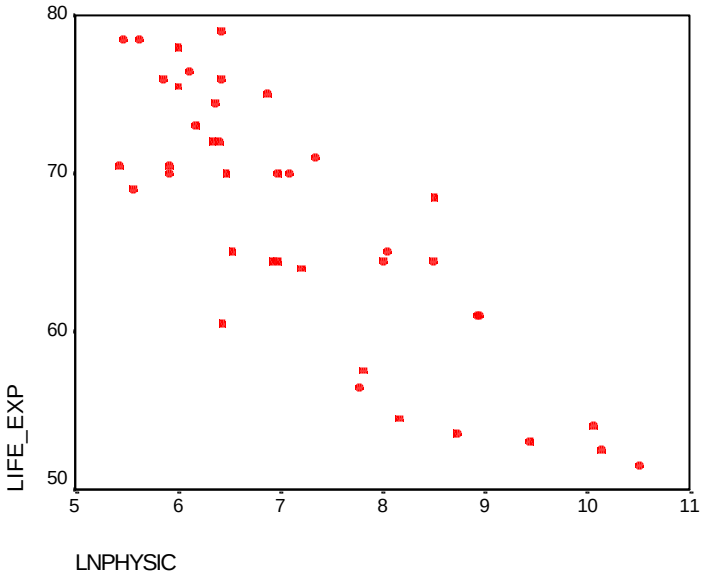
Analisi

I dati si possono rappresentare graficamente per mezzo di un diagramma di dispersione.
Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple**. Scegliamo come **Y-axis** la variabile LIFE_EXP e come **X-axis** la variabile PHYSIC.



Dal diagramma di dispersione appare evidente che le variabili LIFE_EXP e PHYSIC sono correlate negativamente, ma LIFE_EXP non è una funzione lineare di PHYSIC. L'andamento delle osservazioni fa pensare che una trasformazione logaritmica della variabile PHYSIC può essere opportuna.
Dal menù scegliamo **Trasform**, quindi **Compute.** Come **Target Variable** scegliamo LNPHYSIC, come **Numeric Expression "ln(PHYSIC)"**.

Le variabili LIFE_EXP e LNPHYSIC sono legate da una relazione lineare, come mostra il diagramma di dispersione.



Ipotizziamo che valga il seguente modello:

$$LIFE\_EXP = b_0 + b_1 \cdot LNPHYSIC + e$$

e supponiamo che valgano le ipotesi forti.

Dal menu **Analyse**, selezioniamo **Regression** e quindi **Linear**. Selezioniamo come **Dependent variable** LIFE_EXP e come **Independent(s) variable** GDP. Dalla finestra **Linear Regression** selezioniamo

- **Statistics** => **Descriptives** e dalla finestra **Residuals**, la voce **Casewise Diagnostics**, con **Outliers outside:** 2 **standard deviations.**
- **Save** => dalla finestra **Predicted values** la voce **Unstandardized** e dalla finestra **Residuals** la voce **Standardized** (in questo modo vengono salvate nella **Window SPSS data editor**, contenente la matrice di dati, le variabile PRE_1 e ZRE_1)
- **Plots** => attiviamo **Histogram** e **Normal probability plot**

Analisi dell'output

La tabella **Descriptive Statistics** contiene media e deviazione standard delle variabili prese in esame. La speranza di vita media alla nascita è pari a 67.0375 anni.

Descriptive Statistics

	Mean	Std. Deviation	N
LIFE_EXP	67.0375	8.2488	40
LNPHYSIC	7.2040	1.3800	40

La tabella **Correlations** contiene l'indice di correlazione di Pearson tra le variabili LIFE_EXP e LNPHYSIC (-0.832). E' altamente significativo in quanto caratterizzato da un p-value prossimo a zero, quindi l'ipotesi nulla che la correlazione sia zero è rifiutata.

Correlations

		LIFE_EXP	LNPHYSIC
Pearson Correlation	LIFE_EXP	1.000	-.832
	LNPHYSIC	-.832	1.000
Sig. (1-tailed)	LIFE_EXP	.	.000
	LNPHYSIC	.000	.
N	LIFE_EXP	40	40
	LNPHYSIC	40	40

La tabella **Coefficients** contiene

- le stime dei parametri del modello (intercetta e coefficiente angolare) (B),
- gli errori standard degli stimatori ottenuti con il metodo dei minimi quadrati (Std.Error), le statistiche (t)
- e i *p-values* (Sig.) dei test di Students che verificano se i parametri siano significativamente diversi da zero.

Il p-value del test che verifica H_0: $b_0 = 0$ contro H_1: $b_0 \neq 0$ è prossimo a zero, quindi a tutti i livelli di significatività si rifiuta l'ipotesi che b_0 sia zero.

Analogamente il p-value del test che verifica H_0: $b_1 = 0$ contro H_1: $b_1 \neq 0$ è prossimo a zero, quindi a tutti i livelli di significatività si rifiuta l'ipotesi che b_1 sia zero.

Coefficients[a]

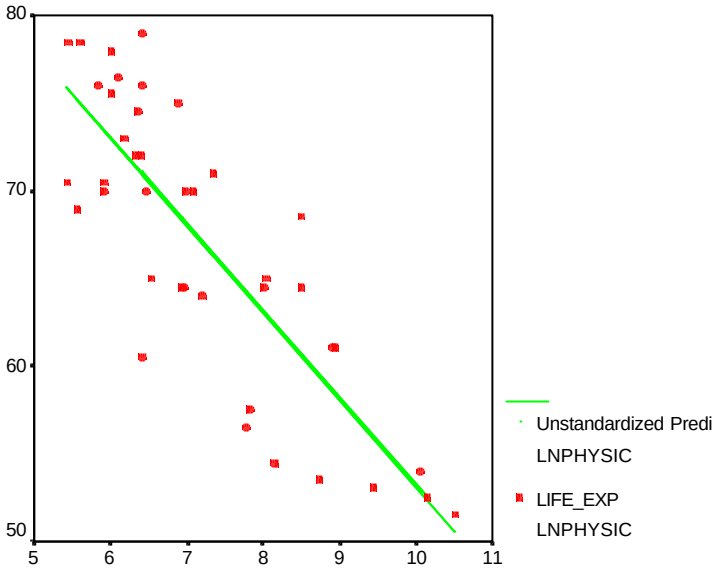
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	102.873	3.942		26.098	.000
	LNPHYSIC	-4.974	.538	-.832	-9.252	.000

a. Dependent Variable: LIFE_EXP

Il modello lineare stimato è

$$LIFE\_EXP = 102.873 - 4.974 \cdot LNPHYSIC$$

Rappresentiamo ora sullo stesso grafico i valori osservati di LIFE_EXP e LNPHYSIC e la retta interpolante (o retta di regressione). Dal menu **Graphs** selezioniamo **Scatter** e quindi **Overlay**. Come **Y-X Pairs** scegliamo dapprima la coppia di variabili LIFE_EXP-LNPHYSIC e successivamente la coppia di variabili PRE_1-LNPHYSIC.



La capacità esplicativa della variabile esplicativa GDP di rappresentare la variabile dipendente LIFE_EXP per mezzo di una retta può essere misurata utilizzando il coefficiente di determinazione R^2 $(0 \le R^2 \le 1)$, che è dato dal rapporto tra la devianza spiegata (o devianza del modello) e devianza totale e rappresenta la proporzione di variabilità totale spiegata dal modello.
Nella tabella **Model Summary** leggiamo il valore di R che rappresenta il coefficiente di correlazione lineare tra le due variabili e il valore del coefficiente di determinazione R^2 che è pari a 0.693.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.832[a]	.693	.684	4.6335

a. Predictors: (Constant), LNPHYSIC

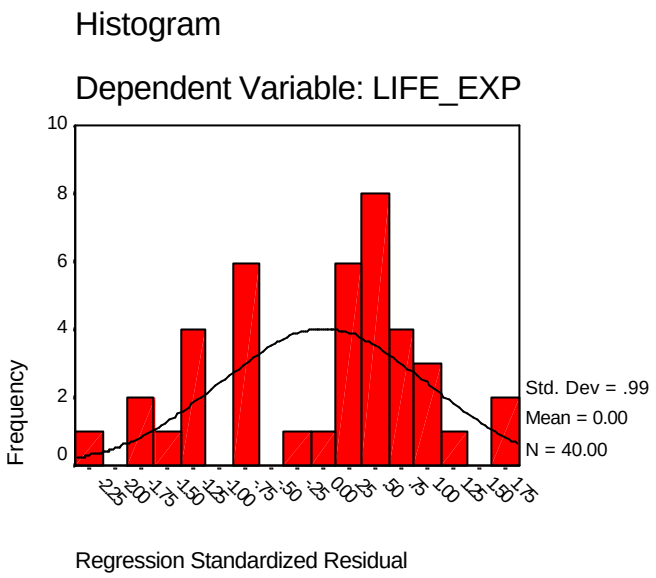
b. Dependent Variable: LIFE_EXP

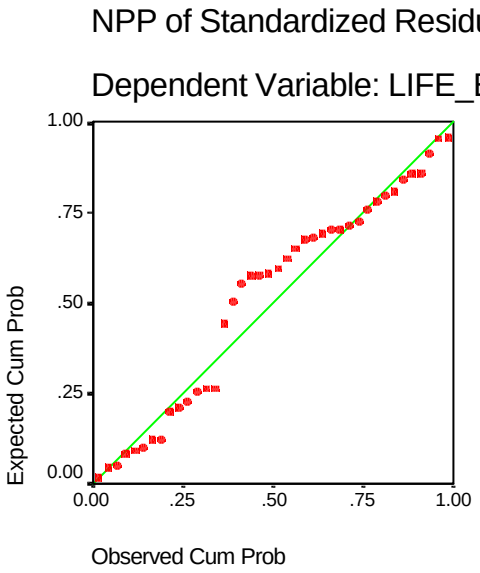
ANALISI DEI RESIDUI

Un ulteriore strumento per controllare la bontà di un modello di regressione è dato dall'analisi dei residui. Se sono verificate le ipotesi forti

- i residui sono normali di media zero e varianza costante;
- i residui sono indipendenti
- i residui e i valori stimati sono indipendenti

I due grafici successivi, un istogramma e un *normal probability plot* (NPP) dei residui standardizzati, sono utilizzati per verificare se sia plausibile l'assunzione di normalità dei residui. Come possiamo osservare la distribuzione dei residui appare leggermente asimmetrica verso sinistra nei dati. Nel NPP, molti punti tendono a disporsi lungo una retta, sebbene vi sia uno scostamento dalla retta tra 0.50 e 0.75 Tenendo conto del numero basso di osservazioni, si può concludere che non c'è sufficiente evidenza di una forte violazione dell'ipotesi di normalità





I

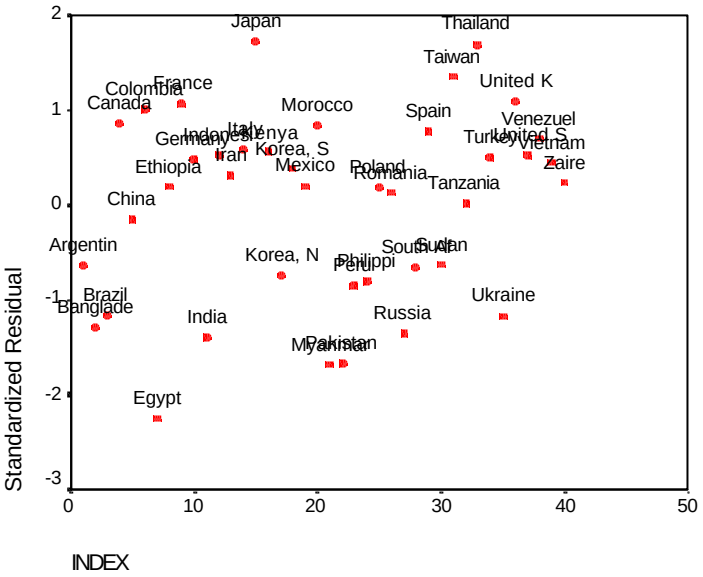
Costruiamo quindi

1. il plot dei residui standardizzati rispetto agli indici del data set.

Costruiamo la variabile INDEX contenente gli indici del data set. Dal menù scegliamo **Trasform**, quindi **Compute.** Come **Target Variable** scegliamo INDEX, come **Numeric Expression *"$casenum"***. Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple**. Come **Y-axis** la variabile ZRE_1 e come **X-axis** la variabile IN DEX.

Questo grafico è utile per rilevare la presenza di possibili outliers, ovvero osservazioni con residui elevati in valore assoluto.

In questo caso, Egitto è l'osservazione con il residuo più elevato in valore assoluto (-2.249, dalla tabella **Casewise Diagnostics**)



Casewise Diagnostics [a]

Case Number	COUNTRY	Std. Residual	LIFE_EXP	Predicted Value	Residual
7	Egypt	-2.249	60.50	70.9215	-10.4215

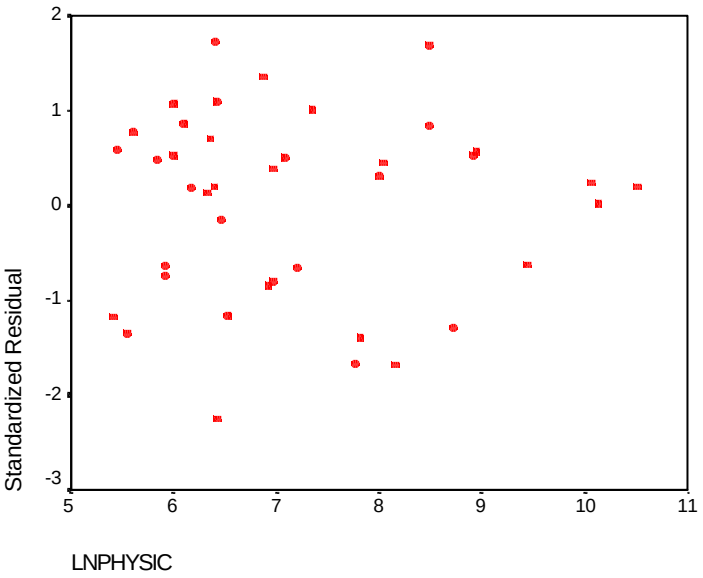
a. Dependent Variable: LIFE_EXP

2. il plot dei residui standardizzati rispetto alla variabile esplicativa LNPHYSIC

Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple**. Come **Y-axis** la variabile ZRE_1 e come **X-axis** la variabile LNPHYSIC.

Questo grafico può mostrare un andamento nei residui che indica non linearità. Inoltre può suggerire la necessità di introdurre un'altra variabile esplicativa.

In questo caso il grafico è soddisfacente, in quanto non rivela alcun particolare andamento.



3. il plot dei residui standardizzati rispetto ai valori stimati.

Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple**. Come **Y-axis** la variabile ZRE_1 e come **X-axis** la variabile PRE_1.

Dal momento che, se sono soddisfatte le ipotesi del modello, i residui e i valori stimati sono indipendenti, nel grafico di punti (PRE_1_i,ZRE_1_i) dovrebbe apparire che i valori di una delle due coordinate non influenzano i valori dell'altra. Questo grafico può anche mostrare se è presente eteroschedasticità, cioè se la varianza dei residui non è costante nel tempo.

In questo caso non c'è evidenza di eteroschedasticità o dipendenza tra i residui e i valori stimati. La nuvola di punti si distribuisce in modo abbastanza casuale.