Esercizio 5
(Scelta delle variabili)

DATI

Il data set macro.sav (o macro.xls) contiene dati relativi al 2001 e riferiti a 24 stati, raccolti dall'UNDP (United Nations Development Programme). Le variabili sono

COUNTRY	name of country
GDP	GDP per capita (PPP US $)
LIFE_EXP	Life expectancy at birth (years)
POP_URB	Urban population (as % of total)
EXP_EDUC	Public expenditure on education as % of GDP (1998-2001)
NET	Net secondary enrolment ratio
TELEPH	Telephone mainlines (per 1000 people)
CELL	Cellular subscribers (per 1000 people)
PAT	Patents granted to residents (per million people)
REC	Receipts of royalties and licence fees (US$ per person)
SCIEN	Scientist and engineers in R & D (per million people)

Domanda 1

Vogliamo studiare quali siano le variabili maggiormente rilevanti per prevedere la speranza di vita (LIFE_EXP).

A questo scopo si vuole costruire un modello di regressione lineare per la speranza di vita, utilizzando il metodo cosiddetto di "backward elimination" per selezionare le variabile esplicative da includere nel modello.

Analisi

Vogliamo costruire e stimare un modello di regressione per la variabile 'speranza di vita' (LIFE_EXP), in funzione delle variabili esplicative disponibile nel data-set.
L'analisi da condurre richiede dunque due passi:

- "selezionare il modello", ovvero studiare quali siano le variabili rilevanti da includere nel modello di regressione;
- procedere alla stima del modello costruito.

SPSS contiene alcune procedure di 'selezione' automatica del modello (backward, forward, stepwise). Qui illustriamo l'uso della procedura 'backward'.
Dal menu **Analyse**, selezioniamo **Regression** e quindi **Linear**.
Selezioniamo come **Dependent variable** LIFE_EXP, come **Independent(s) variable** GDP, POP_URB, EXP_EDUC, NET, TELEPH, CELL, PAT, REC, SCIEN e come metodo di selezione del modello **Method Backward**.
Dalla finestra **Linear Regression** selezioniamo

- **Statistics** => **Descriptives** e **Collinearity diagnostics**

Output

La tabella **Descriptive Statistics** contiene media e deviazione standard delle variabili prese in esame.

Descriptive Statistics

	Mean	Std. Deviation	N
LIFE_EXP	77,2500	2,3969	24
GDP	21677,50	7179,3673	24
POP_URB	77,2708	10,8447	24
EXP_EDUC	5,1208	1,1108	24
NET	88,5833	8,7919	24
TELEPH	494,7500	158,7117	24
CELL	623,5833	206,8269	24
PATENTS	169,3333	269,5176	24
RECEIPTS	47,9042	50,5916	24
SCIENTIS	2505,3750	1361,2545	24

Dalla tabella **Correlations**, riportata sotto, si osserva che la variabile che ha correlazione (positiva) più alta con la variabile dipendente LIFE_EXP è NET (0.732), mentre la variabile che ha correlazione (positiva) più bassa con la variabile dipendente LIFE_EXP è PATENTS (0.292). Al di fuori della prima riga e della prima colonna vi sono le correlazioni tra le variabili indipendenti. Possiamo osservare come alcune variabili siano fortemente correlate tra loro (ad esempio NET, TELEPH , GDP, SCIENTIS), quindi ci aspettiamo che un modello che include tutte le variabili disponibili come regressori risulti sovraparametrizzato.

Correlations

		LIFE_EXP	GDP	POP_URB	EXP_EDUC	NET	TELEPH	CELL	PATENTS	RECEIPTS	SCIENTIS
Pearson Correlation	LIFE_EXP	1.000	.689	.327	.347	.732	.722	.566	.292	.445	.620
	GDP	.689	1.000	-.028	.257	.720	.840	.550	.150	.613	.687
	POP_URB	.327	-.028	1.000	.026	.059	.180	-.095	.170	.136	-.009
	EXP_EDUC	.347	.257	.026	1.000	.289	.435	.482	-.295	.310	.296
	NET	.732	.720	.059	.289	1.000	.821	.618	.357	.489	.702
	TELEPH	.722	.840	.180	.435	.821	1.000	.544	.303	.658	.820
	CELL	.566	.550	-.095	.482	.618	.544	1.000	.072	.362	.357
	PATENTS	.292	.150	.170	-.295	.357	.303	.072	1.000	.199	.435
	RECEIPTS	.445	.613	.136	.310	.489	.658	.362	.199	1.000	.667
	SCIENTIS	.620	.687	-.009	.296	.702	.820	.357	.435	.667	1.000

Si tratta dunque di studiare quali siano le variabili maggiormente rilevanti per prevedere la speranza di vita.

La tabella Variables Entered/Removed contiene i risultati del metodo di selezione delle variabili detto backward elimination. SPSS stima un modello con tutti e nove i regressori. Quindi ad ogni *step* elimina la variabile meno significativa. Nella colonna variables removed, PATENTS è la prima variabile eliminata e così via fino a che nel modello rimangono le tre variabili GDP, POP_URB e CELL.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SCIENTIS, POP_URB, EXP_EDUC, CELL, PATENTS, RECEIPTS, GDP, NET, TELEPH[a]	.	Enter
2	.	PATENTS	Backward (criterion: Probability of F-to-remove >= .100).
3	.	EXP_EDUC	Backward (criterion: Probability of F-to-remove >= .100).
4	.	TELEPH	Backward (criterion: Probability of F-to-remove >= .100).
5	.	NET	Backward (criterion: Probability of F-to-remove >= .100).
6	.	RECEIPTS	Backward (criterion: Probability of F-to-remove >= .100).
7	.	SCIENTIS	Backward (criterion: Probability of F-to-remove >= .100).

a. All requested variables entered.

b. Dependent Variable: LIFE_EXP

Il valore di R^2 per il modello con tutti e nove predittori (modello "pieno") è 0.773, e per l'ultimo modello indicato, con tre predittori, è minore, pari a 0.662. Il confronto fra modelli va fatto tuttavia sulla base dell'indice "aggiustato", che introduce una penalità per il maggior numero di parametri. Il valore di R^2_{ad} per il modello pieno è 0.628, mentre per il modello a 3 regressori R^2_{ad} è 0.611. La differenza fra i valori di R^2_{ad} nei due modelli è ridotta.

R^2_{ad} assume il valore più elevato (0.654) in corrispondenza al quinto modello, caratterizzato da cinque regressori: POP_URB, GDP, CELL, SCIENTIS e RECEIPTS.

Model Summary[h]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.879[a]	.773	.628	1.4625
2	.877[b]	.770	.647	1.4236
3	.870[c]	.757	.651	1.4156
4	.861[d]	.741	.650	1.4176
5	.854[e]	.730	.654	1.4090
6	.839[f]	.704	.642	1.4348
7	.814[g]	.662	.611	1.4945

a. Predictors: (Constant), SCIENTIS, POP_URB, EXP_EDUC, CELL, PATENTS, RECEIPTS, GDP, NET, TELEPH

b. Predictors: (Constant), SCIENTIS, POP_URB, EXP_EDUC, CELL, RECEIPTS, GDP, NET, TELEPH

c. Predictors: (Constant), SCIENTIS, POP_URB, CELL, RECEIPTS, GDP, NET, TELEPH

d. Predictors: (Constant), SCIENTIS, POP_URB, CELL, RECEIPTS, GDP, NET

e. Predictors: (Constant), SCIENTIS, POP_URB, CELL, RECEIPTS, GDP

f. Predictors: (Constant), SCIENTIS, POP_URB, CELL, GDP

g. Predictors: (Constant), POP_URB, CELL, GDP

h. Dependent Variable: LIFE_EXP

La tabella seguente contiene la parte della tabella **Coefficients** relativa alla stima dei coefficienti del modello completo. Il modello completo è caratterizzato da sette valori della

statistica T molto bassi (minori di 2 in valore assoluto). Le variabili corrispondenti sono perciò candidate ad essere eliminate. Il metodo Backward elimination elimina inizialmente proprio la variabile con il valore della statistica T più basso (PATENTS).

In questa tabella compaiono anche le Collinearity Statistics: Tolerance e VIF**.**

Per la variabile i-esima la statistica Tolerance è data da

$$\text{Tolerance}=1-R_i^2$$

dove R_i^2 è il coefficiente di correlazione multipla tra quella variabile e le altre variabili indipendenti. I valori di questa statistica sono compresi tra 0 e 1. Quando questa statistica assume valori piccoli, allora la variabile è una combinazione lineare delle altre variabili indipendenti. In SPSS il valore soglia per questa statistica è 0.0001. Per essere inclusa nel modello, una variabile deve essere caratterizzata da un valore della statistica Tolerance maggiore del valore soglia, indipendentemente dal metodo di selezione usato.
La statistica VIF (Variance Inflation Factor**)** è il reciproco della statistica Tolerance.

In questo caso la variabile che ha tolleranza più bassa e quindi VIF più alta, è TELEPH. Ciò significa che il 91.1% della variabilità di TELEPH, nel modello 1, è spiegato dagli altri predittori, segnalando quindi la presenza di forte collinerità. La variabile TELEPH è fra quelle che saranno escluse nella costruzione del modello 'ottimale'.

La tabella successiva contiene la parte della tabella **Coefficients** relativa alla stima dei coefficienti per i modelli 2, 3, 4 , 5, 6, 7. Ad ogni passo viene eliminata la variabile con il valore più basso della statistica T. Il valore più basso della statistica T nel modello 7 è 1.990. Dal momento che, per *default*, vengono escluse le variabili che hanno una statistica T associata minore di 1.646, non viene eliminata alcuna altra variabile.
Formalmente, per eliminare le variabili non significative SPSS fa riferimento, in questo caso, al *p-value* relativo alla statistica F-to-remove che è data dal quadrato della statistica T. Per *default* le variabili con p-value minore di 0.10 vengono eliminate dal modello.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	56.055	5.724		9.793	.000		
	GDP	2.03E-04	.000	.608	2.058	.059	.186	5.386
	POP_URB	.101	.034	.457	2.944	.011	.672	1.487
	EXP_EDUC	.430	.435	.199	.987	.340	.398	2.511
	NET	9.16E-02	.070	.336	1.306	.213	.245	4.089
	TELEPH	-8.5E-03	.006	-.565	-1.326	.206	.089	11.211
	CELL	2.36E-03	.002	.204	1.025	.323	.410	2.436
	PATENTS	7.60E-04	.002	.085	.461	.652	.470	2.127
	RECEIPTS	-9.5E-03	.009	-.200	-1.066	.304	.461	2.167
	SCIENTIS	7.00E-04	.001	.398	1.392	.186	.198	5.038

a. Dependent Variable: LIFE_EXP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
2	(Constant)	56.121	5.570		10.075	.000		
	POP_URB	.103	.033	.468	3.141	.007	.690	1.449
	EXP_EDUC	.316	.349	.146	.906	.379	.587	1.703
	NET	9.40E-02	.068	.345	1.381	.188	.246	4.067
	TELEPH	-8.1E-03	.006	-.534	-1.304	.212	.091	10.935
	CELL	2.60E-03	.002	.225	1.194	.251	.433	2.307
	RECEIPTS	-9.6E-03	.009	-.203	-1.116	.282	.462	2.163
	SCIENTIS	7.89E-04	.000	.448	1.746	.101	.233	4.295
	GDP	1.86E-04	.000	.558	2.085	.055	.214	4.672
3	(Constant)	58.511	4.878		11.995	.000		
	POP_URB	9.95E-02	.032	.450	3.064	.007	.703	1.422
	NET	7.74E-02	.065	.284	1.187	.253	.265	3.773
	TELEPH	-5.7E-03	.006	-.379	-1.025	.321	.111	9.030
	CELL	3.41E-03	.002	.294	1.723	.104	.520	1.922
	RECEIPTS	-9.0E-03	.009	-.190	-1.053	.308	.465	2.149
	SCIENTIS	7.59E-04	.000	.431	1.693	.110	.234	4.272
	GDP	1.58E-04	.000	.474	1.899	.076	.243	4.114
4	(Constant)	60.815	4.335		14.028	.000		
	POP_URB	8.39E-02	.029	.380	2.920	.010	.899	1.112
	NET	5.41E-02	.061	.199	.884	.389	.302	3.315
	CELL	3.07E-03	.002	.265	1.572	.134	.535	1.868
	RECEIPTS	-9.4E-03	.009	-.198	-1.098	.288	.466	2.145
	SCIENTIS	5.18E-04	.000	.294	1.356	.193	.322	3.101
	GDP	1.10E-04	.000	.330	1.598	.129	.356	2.809
5	(Constant)	63.900	2.558		24.977	.000		
	POP_URB	8.95E-02	.028	.405	3.210	.005	.944	1.059
	CELL	3.88E-03	.002	.335	2.263	.036	.686	1.457
	RECEIPTS	-1.1E-02	.008	-.230	-1.304	.209	.485	2.063
	SCIENTIS	6.86E-04	.000	.389	2.075	.053	.427	2.344
	GDP	1.30E-04	.000	.389	2.004	.060	.398	2.514
6	(Constant)	64.999	2.460		26.426	.000		
	POP_URB	8.17E-02	.028	.370	2.946	.008	.990	1.010
	CELL	3.70E-03	.002	.319	2.127	.047	.691	1.448
	SCIENTIS	4.98E-04	.000	.283	1.643	.117	.527	1.898
	GDP	1.10E-04	.000	.330	1.714	.103	.421	2.374
7	(Constant)	64.832	2.560		25.325	.000		
	POP_URB	8.21E-02	.029	.372	2.843	.010	.990	1.010
	CELL	3.61E-03	.002	.311	1.990	.060	.691	1.446
	GDP	1.76E-04	.000	.528	3.393	.003	.697	1.434

a. Dependent Variable: LIFE_EXP

Al fine di verificare se si è in presenza di un problema di **multicollinearità**, SPSS calcola anche

- Gli autovalori della matrice $X_{(s)}'X_{(s)}$ (Eigenvalue), dove X è la matrice disegno di dimensione $(n\times(p+1))$, con n numero di osservazioni e p numero di regressori, e $X_{(s)} = XD_{(s)}^{(-1)}$, con $D_{(s)} = diag\left(\left\|x_{[0]}\right\|,\ldots,\left\|x_{[k]}\right\|\right)$ e $x_{[j]}$ j-esimo vettore colonna della matrice X.
- i condition indices, dati da

$$\text{condition index}_j = \sqrt{l_{\max}/l_j}$$

dove l_j è il j-esimo autovalore della matrice $X_{(s)}'X_{(s)}$ e $l_{\max} = \max_{0\le j\le p} l_j$

Un condition index molto grande (maggiore di 30) indica un elevato grado di collinearità

- la matrice coefficient variance-decomposition. Ogni riga di tale matrice mostra la proporzione di varianza dello stimatore di ogni coefficiente di regressione attribuibile all'autovalore corrispondente.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions									
				(Constant)	GDP	POP_URB	EXP_EDUC	NET	TELEPH	CELL	PATENTS	RECEIPTS	SCIENTIS
1	1	8,637	1,000	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
	2	,692	3,534	,00	,00	,00	,00	,00	,00	,00	,40	,01	,00
	3	,438	4,441	,00	,00	,00	,00	,00	,00	,00	,07	,38	,00
	4	9,377E-02	9,598	,00	,02	,02	,00	,00	,01	,00	,05	,48	,20
	5	6,575E-02	11,461	,00	,00	,02	,00	,00	,00	,41	,04	,00	,09
	6	3,883E-02	14,915	,00	,22	,00	,13	,00	,01	,05	,00	,00	,13
	7	1,658E-02	22,822	,03	,00	,00	,22	,01	,18	,11	,10	,02	,27
	8	1,133E-02	27,610	,02	,23	,36	,37	,00	,07	,26	,30	,02	,10
	9	5,367E-03	40,117	,03	,47	,50	,14	,11	,39	,04	,04	,08	,20
	10	1,266E-03	82,614	,93	,05	,09	,13	,88	,35	,14	,00	,00	,00
2	1	8,279	1,000	,00	,00	,00	,00	,00	,00	,00		,00	,00
	2	,464	4,224	,00	,00	,00	,00	,00	,00	,00		,36	,01
	3	,105	8,899	,00	,01	,01	,01	,00	,00	,00		,50	,26
	4	6,989E-02	10,884	,00	,01	,03	,00	,00	,00	,37		,01	,04
	5	3,902E-02	14,567	,00	,23	,01	,23	,00	,01	,03		,00	,16
	6	2,233E-02	19,255	,00	,11	,03	,47	,00	,03	,33		,02	,13
	7	1,403E-02	24,291	,04	,15	,20	,03	,01	,24	,04		,03	,04
	8	5,530E-03	38,694	,03	,42	,62	,08	,10	,37	,08		,09	,36
	9	1,266E-03	80,882	,93	,06	,10	,18	,89	,35	,14		,00	,00
3	1	7,329	1,000	,00	,00	,00		,00	,00	,00		,00	,00
	2	,446	4,052	,00	,00	,00		,00	,00	,00		,37	,01
	3	,101	8,511	,00	,01	,01		,00	,00	,00		,49	,28
	4	6,982E-02	10,246	,00	,02	,03		,00	,00	,45		,01	,05
	5	3,152E-02	15,249	,00	,44	,00		,00	,02	,28		,01	,24
	6	1,427E-02	22,663	,05	,15	,15		,01	,32	,01		,03	,06
	7	5,995E-03	34,966	,04	,38	,73		,07	,42	,18		,07	,36
	8	1,521E-03	69,420	,90	,01	,06		,92	,23	,07		,01	,01
4	1	6,347	1,000	,00	,00	,00		,00		,00		,00	,00
	2	,446	3,771	,00	,00	,00		,00		,00		,38	,01
	3	9,623E-02	8,121	,00	,01	,02		,00		,00		,47	,48
	4	6,972E-02	9,541	,00	,03	,04		,00		,48		,01	,05
	5	2,996E-02	14,556	,00	,84	,00		,00		,27		,01	,20
	6	8,939E-03	26,646	,16	,05	,94		,05		,09		,11	,04
	7	1,873E-03	58,213	,84	,07	,00		,95		,17		,02	,22
5	1	5,378	1,000	,00	,00	,00				,00		,01	,00
	2	,419	3,581	,00	,00	,00				,01		,41	,01
	3	9,596E-02	7,486	,00	,01	,02				,01		,47	,64
	4	6,869E-02	8,849	,01	,02	,05				,59		,01	,07
	5	2,993E-02	13,406	,00	,93	,01				,35		,01	,27
	6	7,596E-03	26,610	,98	,03	,92				,05		,09	,01
6	1	4,728	1,000	,00	,00	,00				,00			,00
	2	,164	5,363	,01	,01	,02				,01			,44
	3	6,898E-02	8,279	,02	,02	,05				,59			,14
	4	3,016E-02	12,520	,00	,96	,00				,35			,41
	5	8,343E-03	23,805	,97	,01	,93				,05			,00
7	1	3,865	1,000	,00	,00	,00				,00			
	2	8,381E-02	6,791	,03	,18	,07				,21			
	3	4,290E-02	9,492	,00	,81	,00				,73			
	4	8,356E-03	21,507	,97	,01	,93				,05			

a. Dependent Variable: LIFE_EXP

L'ultima riga della matrice coefficient variance-decomposition, relativa al settimo modello, mostra che la proporzione di varianza dello stimatore dell'intercetta e dello stimatore del coefficiente di POP_URB, attribuibili al quarto autovalore, sono molto elevate, anche se il condition index rimane al di sotto del valore di soglia.

Il modello finale, stimato con il metodo Backward elimination, è

$$LIFE\_EXP = 64.832 + 0.000176 \cdot GDP + 0.0821 \cdot POP\_URB + 0.00361 \cdot CELL$$

Esaminando i Beta coefficients (si tratta delle stime dei coefficienti di regressione basate su variabili standardizzate e quindi espresse nella stessa unità di misura), nella tabella **Coefficients,** si può osservare che GDP, POP_URB e CELL rispettivamente, hanno un peso via via crescente nella previsione di LIFE_EXP.

Correlazione parziale e analisi dei residui.

Approfondiamo ora l'analisi del modello finale ottenuto con la procedura di 'backward selection'. Per poter svolgere ulteriori elaborazioni, ristimiamolo separatamente.

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

- **Statistics** => **Part and Partial correlations** e dalla finestra **Residuals** => **Casewise Diagnostics**, con **Outliers outside:** 2 **standard deviations.**
- **Save** => dalla finestra **Predicted values** => **Unstandardized**
  dalla finestra **Residuals** => **Standardized** (in questo modo vengono salvate nella **Window SPSS data editor**, contenente la matrice di dat,i le variabili PRE_1 (che contiene i valori stimati) e la variabile ZRE_1 (che contiene i residui standardizzati) e **Studentized deleted**
  dalla finestra **Distances** => **Cook's** and **Leverage values**
  dalla finestra **Influence Statistics** => **Standardized DfBeta(s)** e **DfFit**
- **Plots** => **Normal probability plot** e **Produce all partial plots**

I partial residuals plots, riportati sotto, forniscono una versione grafica della **correlazione parziale** di ogni variabile esplicativa con la variabile dipendente, dopo che è stata rimossa l'influenza delle altre variabili. I valori assunti dai coefficienti di correlazione parziale sono contenuti in una parte della tabella **Coefficients**, riportata di seguito

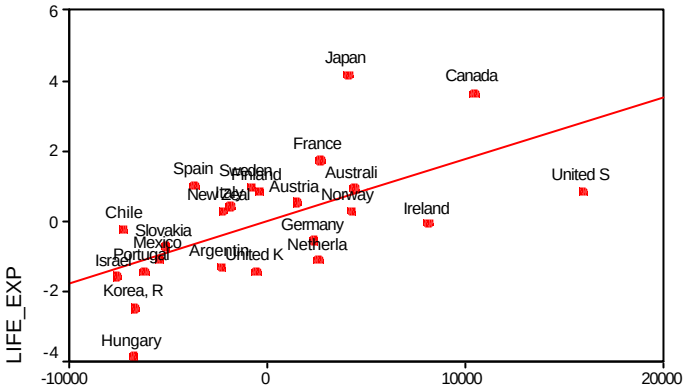
Coefficients[a]

Model		Correlations	
		Zero-order	Partial
1	GDP	,689	,604
	POP_URB	,327	,537
	CELL	,566	,407

a. Dependent Variable: LIFE_EXP

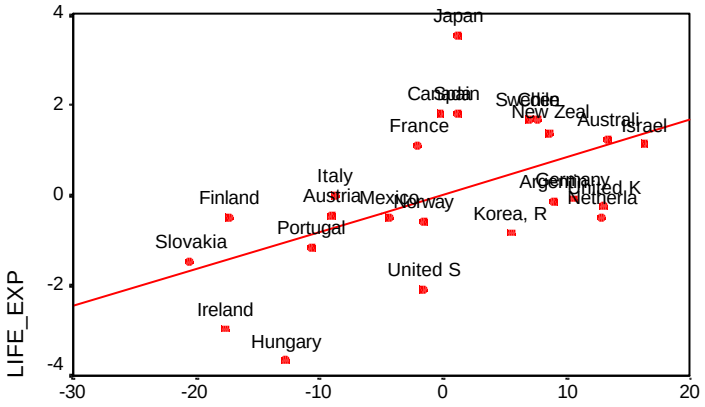
Partial Regression Plot

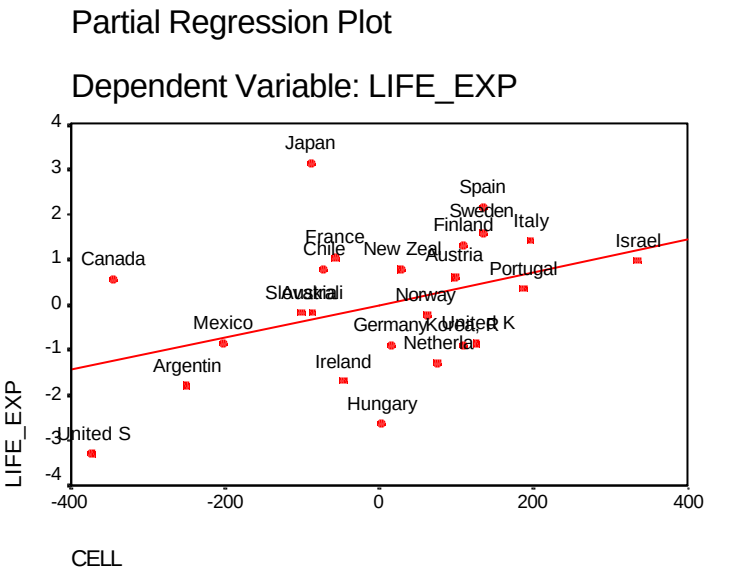
Dependent Variable: LIFE_EXP



Partial Regression Plot

Dependent Variable: LIFE_EXP

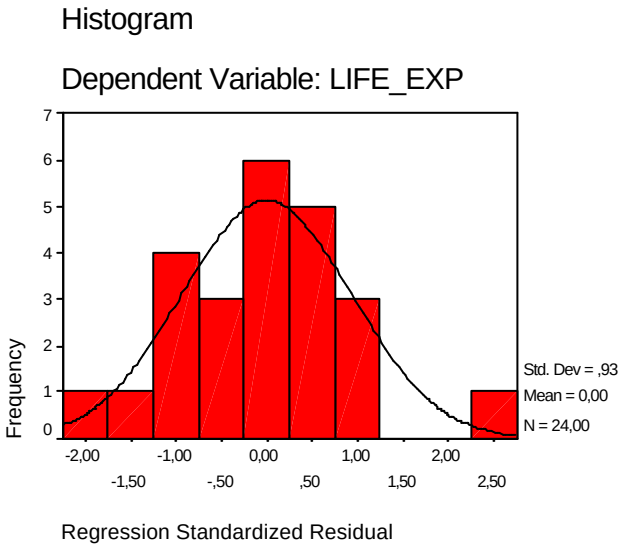


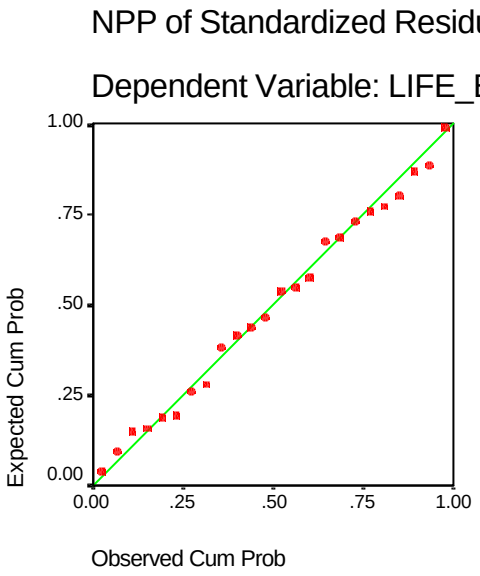


Uno strumento per controllare la bontà di un modello di regressione è dato dall'**analisi dei residui.** Se sono verificate le ipotesi forti, allora

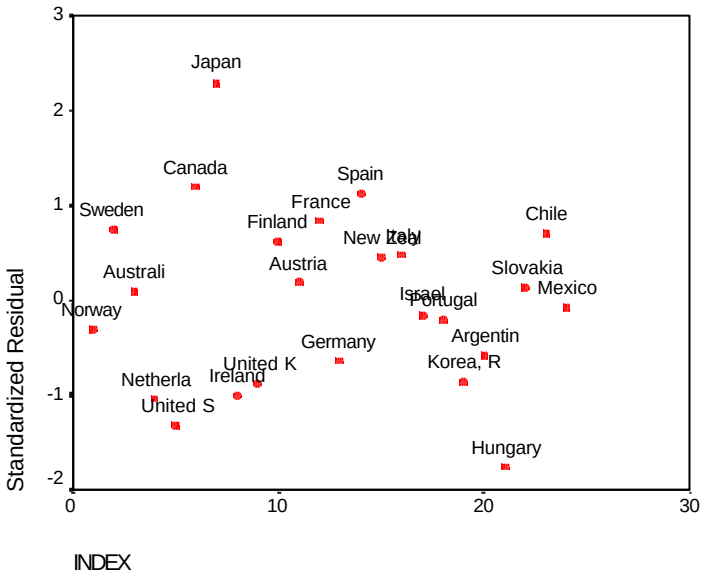
- i residui sono normali di media zero e varianza costante;
- i residui sono indipendenti
- i residui e i valori stimati sono indipendenti

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 i residui seguono approssimativamente una distribuzione normale, sebbene sia riscontrabile una certa asimmetria nei dati. Nel NPP, i punti tendono a disporsi approssimativamente lungo una retta. Si può concludere che i residui standardizzati sono realizzazioni di una distribuzione normale standard.





Il plot dei residui standardizzati rispetto agli indici del data set mostra che l'osservazione con il residuo più elevato (2.299, si veda la tabella **Casewise diagnostic**, dove sono riportati i residui standardizzati maggiori di 2 in valore assoluto) è quella relativa al Giappone.

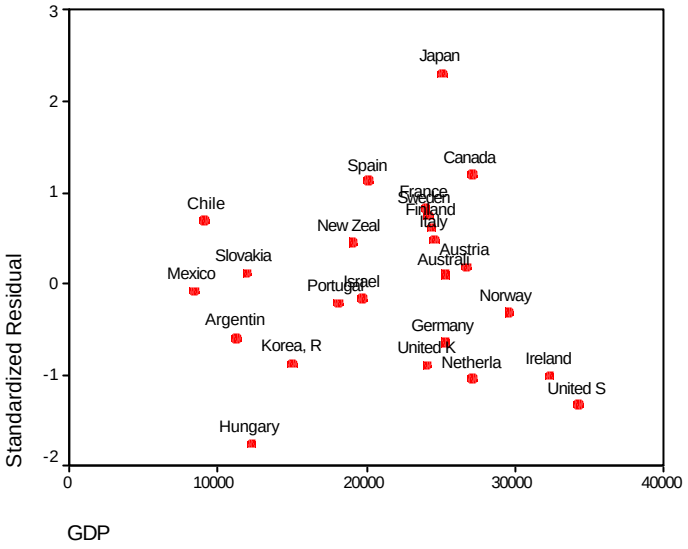


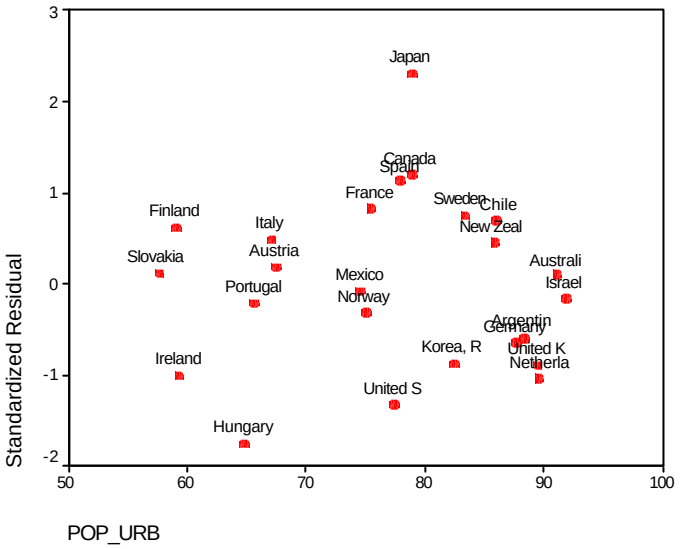
Casewise Diagnostics[a]

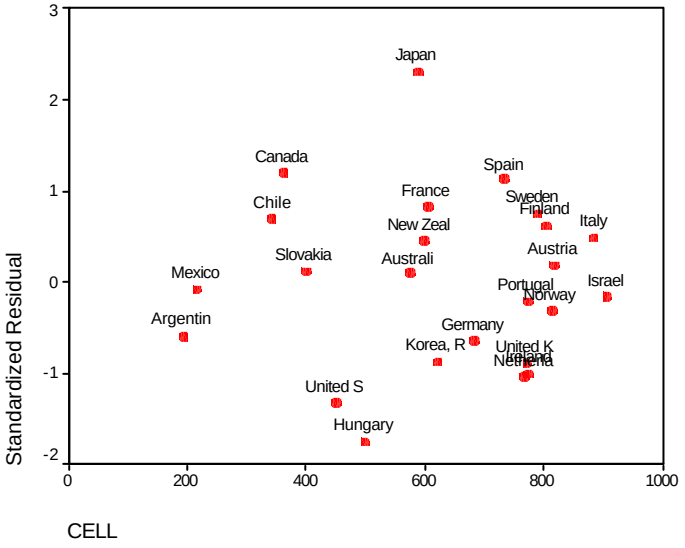
Case Number	Std. Residual	LIFE_EXP	Predicted Value	Residual
7	2,299	81,30	77,8645	3,4355

a. Dependent Variable: LIFE_EXP

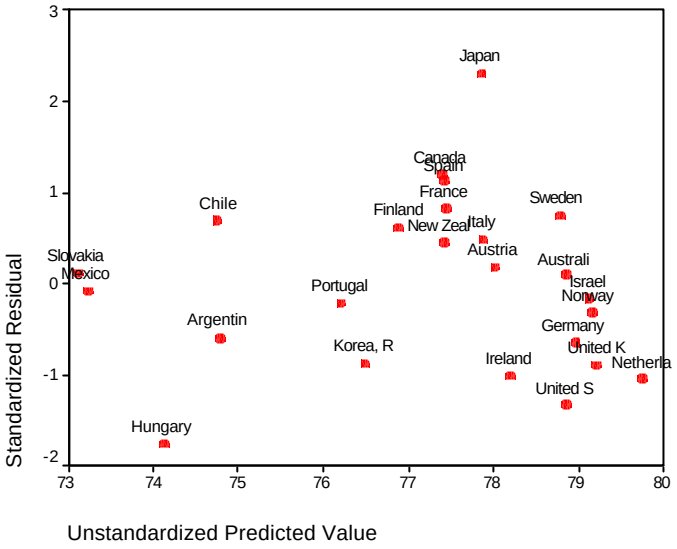
Sono di seguito riportati i grafici dei residui standardizzati rispetto alle variabili esplicative. Tali grafici non mostrano in maniera evidente la presenza di non linearità







E' riportato di seguito il grafico dei residui standardizzati rispetto ai valori stimati. Non c'è forte evidenza di eteroschedasticità o di dipendenza tra residui e valori stimati.



Studiamo ora se vi siano **punti influenti**.

La Cook's distance, calcolata in corrispondenza ad ogni osservazione e salvata come variabile COO_1, può evidenziare la presenza di eventuali punti influenti, cioè di punti che influenzano le stime dei coefficienti di regressione[1].

La statistica Centered Leverage Value, calcolata in corrispondenza ad ogni osservazione e salvata come variabile LEV_1, può evidenziare la presenza di potenziali punti outliers per le variabili indipendenti.[2]

Nel nostro caso entrambe le statistiche assumono valori inferiori al valore di soglia, in corrispondenza ad ogni osservazione.

[1] La distanza di Cook misura l'influenza di un singolo caso sulla stima dei coefficienti di regressione, quando il singolo caso viene rimosso dal processo di stima. Un valore della distanza di Cook >1 indica che il punto è influente

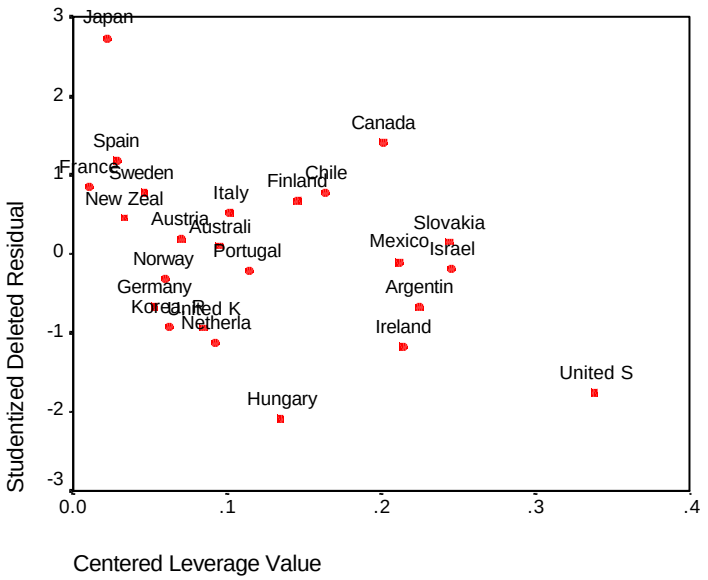
[2] Il leverage di un caso è una misura della distanza tra il vettore contenente i valori delle variabili esplicative associate a quel caso e la media dei vettori contenenti i valori delle variabili esplicative associate a tutti i casi. Per n>50, p>10 (dove n è il numero dei dati e p è il numero di variabili esplicative) il valore soglia per individuare potenziali punti outlier per le variabili esplicative è 2p/n (Belsley et al., 1980), altrimenti Vellmann e Welsch (1981) suggeriscono 3p/n. Nel nostro caso 3p/n=0.375

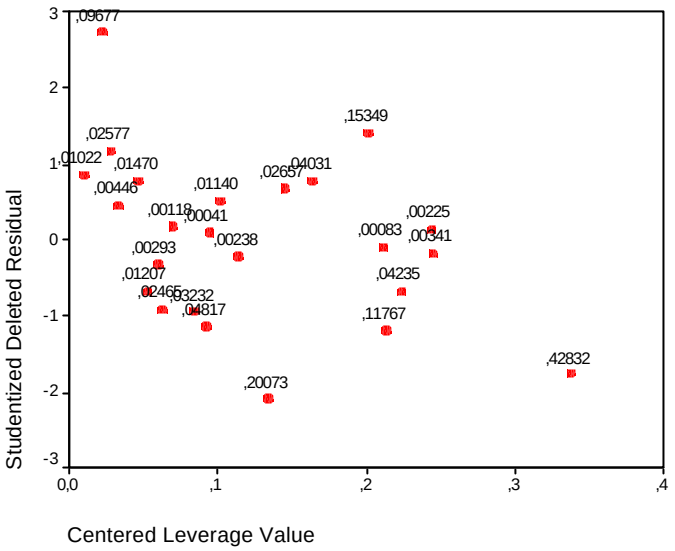
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	73,1106	79,7520	77,2500	1,9501	24
Std. Predicted Value	-2,123	1,283	,000	1,000	24
Standard Error of Predicted Value	,3412	,9211	,5903	,1577	24
Adjusted Predicted Value	73,0349	80,0740	77,3177	2,0125	24
Residual	-2,6259	3,4355	-1,24E-14	1,3937	24
Std. Residual	-1,757	2,299	,000	,933	24
Stud. Residual	-1,936	2,376	-,020	1,020	24
Deleted Residual	-3,1884	3,6710	-6,77E-02	1,6812	24
Stud. Deleted Residual	-2,093	2,734	-,015	1,078	24
Mahal. Distance	,240	7,779	2,875	2,007	24
Cook's Distance	,000	,428	,054	,095	24
Centered Leverage Value	,010	,338	,125	,087	24

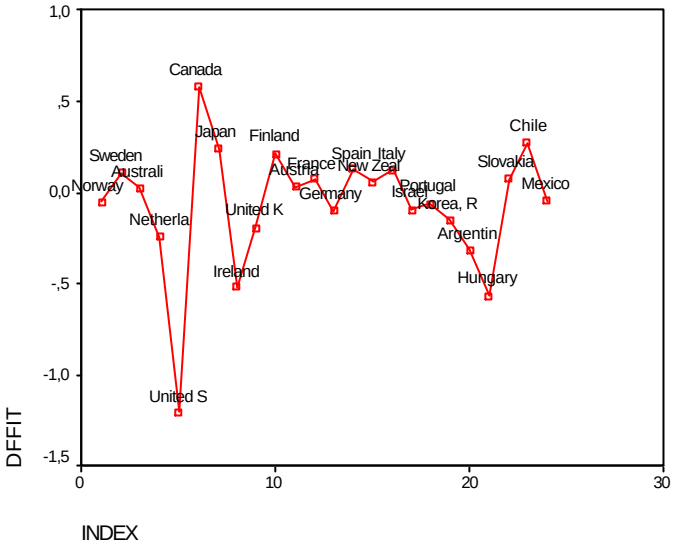
a. Dependent Variable: LIFE_EXP

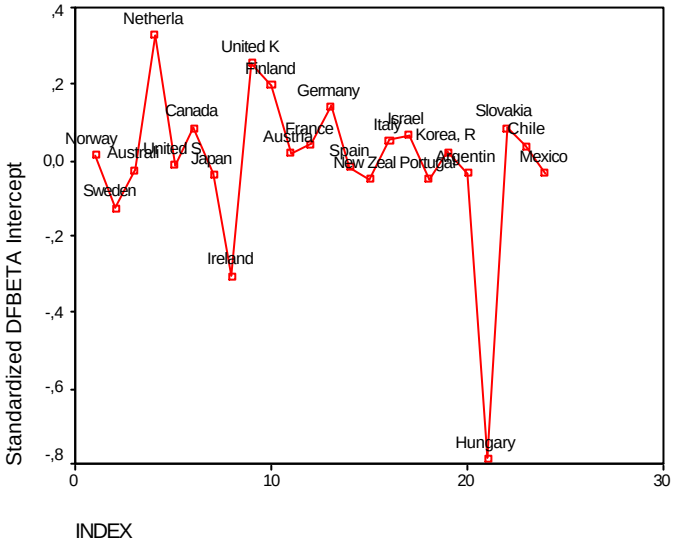
Tuttavia il grafico che riporta i valori della statistica Centered Leverage Value sull'asse delle ascisse e i Studentized Deleted Residuals sull'asse delle ordinate, evidenzia come Stati Uniti sia potenzialmente un'osservazione *outlier* per le variabili indipendenti.
Inoltre lo stesso grafico, con le etichette rappresentate dai valori della distanza di Cook, evidenzia come Stati Uniti, con distanza di Cook pari a 0.42832, sia potenzialmente un'osservazione che influenza la stima dei coefficienti. I successivi due valori più elevati della distanza di Cook corrispondono alle osservazioni Ungheria (0.20073) e Irlanda (0.11767).

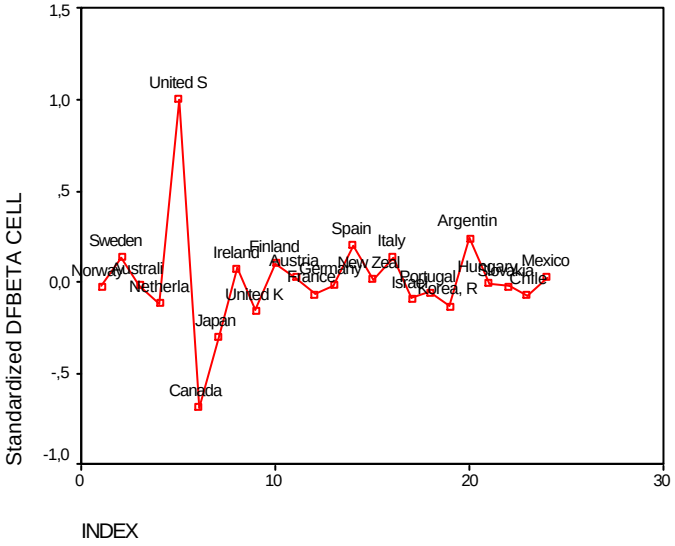


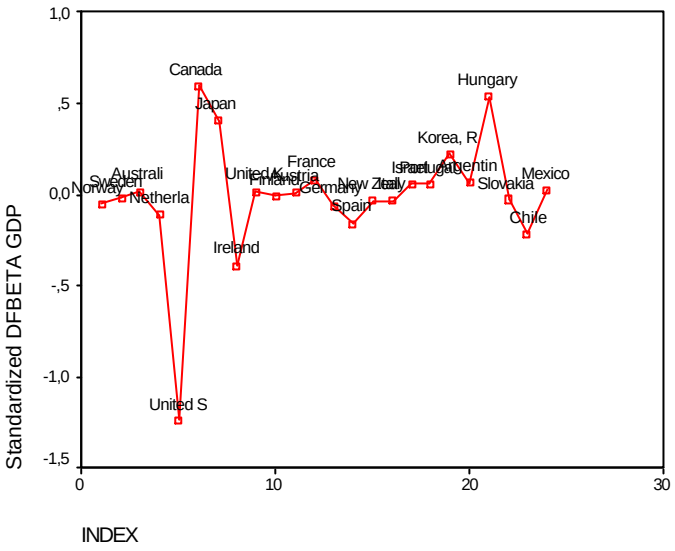


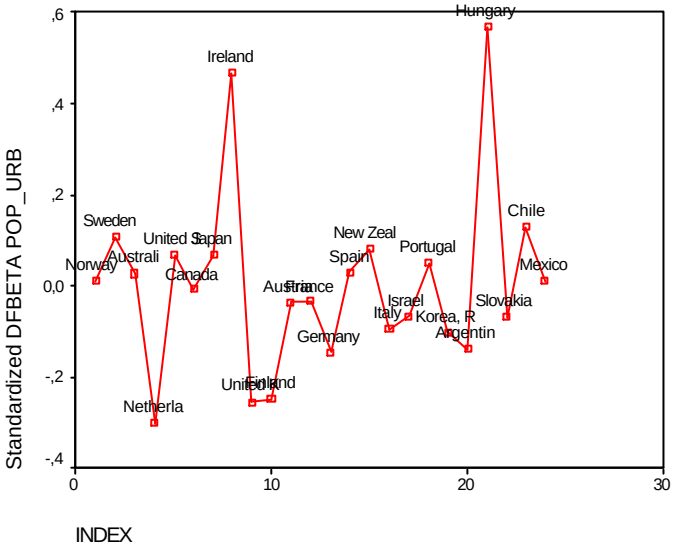
La statistica DFIT di un caso misura l'influenza di quel caso sulla stima dei coefficienti di regressione e sulla loro varianza, quando viene rimosso dal processo di stima . Le statistiche DFBETA(s) di un caso misurano l'influenza di quel caso, quando viene rimosso dal processo di stima, sulle stime di ogni coefficiente di regressione separatamente.


DFFIT
INDEX
1,0
,5
0,0
-,5
-1,0
-1,5
0
10
20
30
Canada
Japan
Finland
Chile
Sweden
Australi
Spain
Italy
Slovakia
Norway
Austria
New Zeal
Mexico
Portugal
Germany
Israel
Korea, R
Netherla
United K
Argentin
Ireland
Hungary
United S


Standardized DFBETA Intercept
INDEX
,4
,2
0,0
-,2
-,4
-,6
-,8
0
10
20
30
Netherla
United K
Finland
Germany
Canada
Israel
Slovakia
Italy
Korea, R
Chile
Norway
France
Austria
Australi
United S
Spain
Argentin
Mexico
Japan
New Zeal
Portugal
Sweden
Ireland
Hungary


Standardized DFBETA CELL
INDEX
1,5
1,0
,5
0,0
-,5
-1,0
0
10
20
30
United S
Argentin
Spain
Sweden
Italy
Ireland
Finland
Austria
Mexico
Norway
Australi
Germany
Hungary
New Zeal
Slovakia
Portugal
Netherla
Israel
Korea, R
Chile
United K
Japan
Canada





I grafici confermano che le osservazioni Stati Uniti, Irlanda e Ungheria sono quelle che influenzano maggiormente la stima dei coefficienti di regressione.