Esempio 1
(Regressione semplice, punti influenti, regressione multipla, multicollinearità)

DATI

Il data set cigarettes.sav (fonte http://www.amstat.org/publications/jse/jse_data_archive.html) contiene dati relativi a 24 marche di sigarette. Le variabili sono

NAME	brand name
TAR	tar content (mg)
NICOT	nicotine content (mg)
WEIGHT	weight in g
CO	Carbon monoxide content (mg)

Domanda 1

*** Studiare la dipendenza del "contenuto di carbonio" (CO) dalle altre variabili presenti nel data set.**

**** In particolare, studiare la dipendenza di CO dal contenuto di catrame (TAR), attraverso un modello di regressione lineare semplice.**

Domanda 2

Ora vogliamo spiegare il contenuto di carbonio CO mediante il contenuto di catrame (TAR) e di nicotina (NICOT) .

Domande:

(a) Costruire un modello di regressione lineare multipla utilizzando come variabile dipendente CO e come variabili esplicative TAR e NICOT.

(b) commentare i risultati ottenuti.

Analisi

* Per avere una prima idea della struttura di dipendenza fra le variabili in esame, possiamo cominciare col costruire la **matrice di correlazione** delle variabili presenti nel data set.

Dal menù **Analyze** => **Correlate** => **Bivariate** =>come **Variables** scegliamo CO, TAR, NICOT, WEIGHT.

L'output è dato da

Correlations

		CO	TAR	NICOT	WEIGHT
CO	Pearson Correlation	1.000	.959**	.926**	.464*
	Sig. (2-tailed)	.	.000	.000	.019
	N	25	25	25	25
TAR	Pearson Correlation	.959**	1.000	.976**	.492*
	Sig. (2-tailed)	.000	.	.000	.012
	N	25	25	25	25
NICOT	Pearson Correlation	.926**	.976**	1.000	.500*
	Sig. (2-tailed)	.000	.000	.	.011
	N	25	25	25	25
WEIGHT	Pearson Correlation	.464*	.492*	.500*	1.000
	Sig. (2-tailed)	.019	.012	.011	.
	N	25	25	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

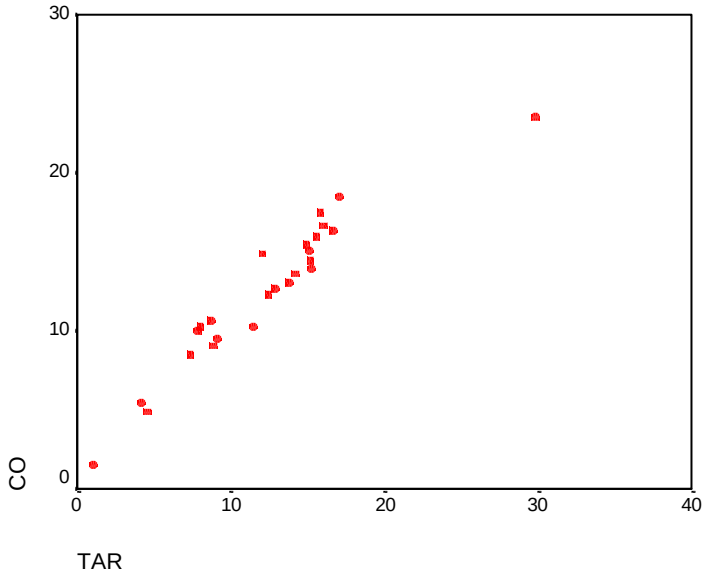
*. Correlation is significant at the 0.05 level (2-tailed).

La variabile CO è fortemente correlata con le variabili TAR (**coefficiente di correlazione lineare** fra CO e TAR =0.959) e NICOT (0.926).

**** Ci proponiamo ora di spiegare CO tramite la variabile TAR, attraverso un modello di regressione lineare.**

E' sempre bene cominciare col rappresentare graficamente i dati per mezzo di un **diagramma di dispersione.**

Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple**. Scegliamo come **Y-axis** la variabile CO e come **X-axis** la variabile TAR.



Dal diagramma di dispersione appare evidente che un modello di regressione lineare è adeguato a rappresentare la relazione tra CO e TAR.
Il grafico mostra anche la presenza di un punto "anomalo" (corrispondente all'osservazione 3), che non rappresenta, tuttavia, una violazione al legame lineare CO e TAR, in quanto appare allineato con la restante nuvola di punti, sebbene isolato rispetto ad essa.

Ipotizziamo che valga il seguente modello:

$$CO = b_0 + b_1 \cdot TAR + e$$

e supponiamo che siano soddisfatte le ipotesi forti.

Dal menu **Analyse**, selezioniamo **Regression** e quindi **Linear**. Selezioniamo come **Dependent variable** CO e come **Independent(s) variable** TAR.

Dalla finestra **Linear Regression** selezioniamo

- **Statistics** => **Descriptives** 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 dati le variabili PRE_1 e ZRE_1 e **Studentized deleted** (viene salvata nella **Window SPSS data editor** la variabile SDR_1)
  dalla finestra **Distances** => **Cook's** and **Leverage values** (vengono salvate nella **Window SPSS data editor** le variabili COO_1 e LEV_1)
  dalla finestra **Influence Statistics** => **Standardized DfBeta(s)** e **DfFit** (vengono salvate nella **Window SPSS data editor** le variabili DIFF_1, SDB0_0, SDB1_1)
- **Plots** => **Histogram** e **Normal probability plot**

Analisi dell'output

La tabella **Descriptive Statistics** contiene **media e deviazione standard** delle variabili prese in esame. Il contenuto medio di monossido di carbonio è pari a 12.528 mg., mentre il contenuto medio di catrame è pari a 12.256 mg.

Descriptive Statistics

	Mean	Std. Deviation	N
CO	12.5280	4.7397	25
TAR	12.2560	5.6861	25

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.

Nella tabella ottenuta, 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.

Anche 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	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.728	.661		4.127	.000
	TAR	.800	.049	.959	16.288	.000

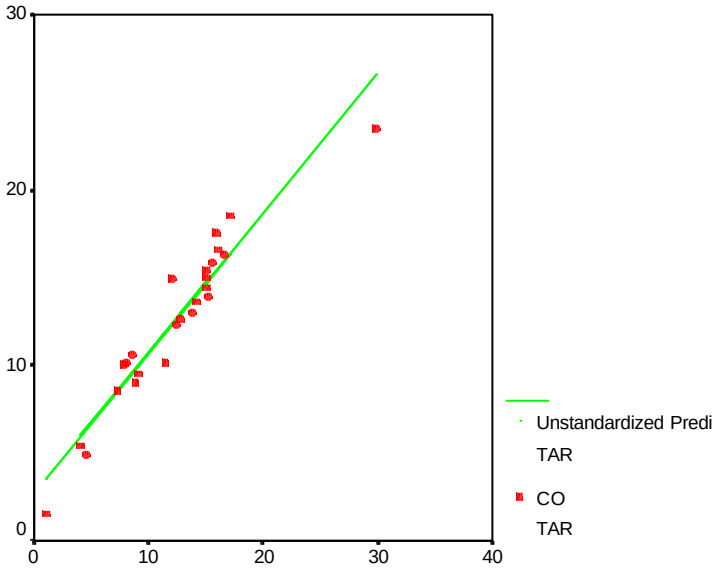
a. Dependent Variable: CO

Il **modello lineare stimato** è

$$CO = 2.728 + 0.8 \cdot TAR$$

All'aumentare del contenuto di catrame di 1 mg , il contenuto di CO aumenta di 0.8 mg.

Rappresentiamo ora sullo stesso grafico i valori osservati di CO e TAR 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 CO-TAR e successivamente la coppia di variabili PRE_1-TAR



La capacità esplicativa della variabile esplicativa TAR di rappresentare la variabile dipendente CO per mezzo di una retta può essere misurata utilizzando il **coefficiente di determinazione** R^2 $(0 \leq R^2 \leq 1)$, 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 0.92. Il modello spiega il 92% della variabilità della variabile CO.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959[a]	.920	.917	1.3675

a. Predictors: (Constant), TAR

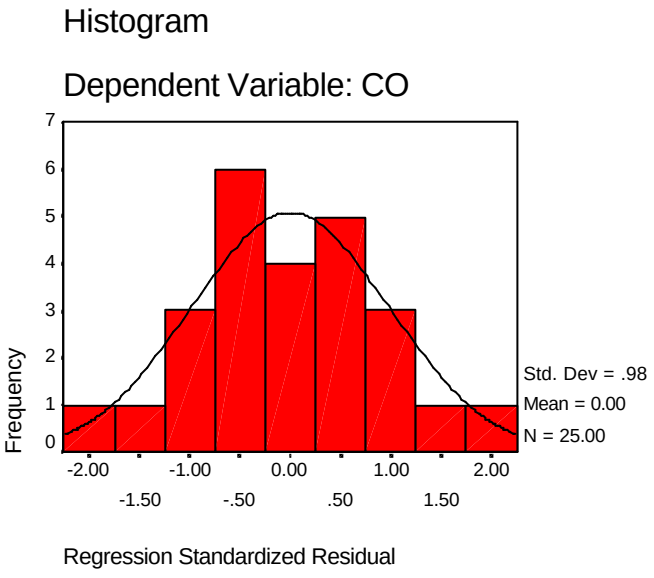
b. Dependent Variable: CO

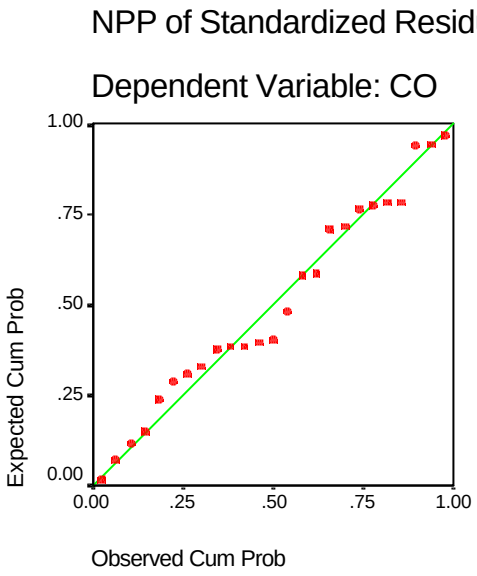
ANALISI DEI RESIDUI

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

- i residui hanno distribuzione normale, con 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 dal grafico ottenuto, i residui seguono approssimativamente una distribuzione normale, sebbene sia riscontrabile una leggera asimmetria nei dati. Nel NPP, molti punti tendono a disporsi lungo una retta. 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

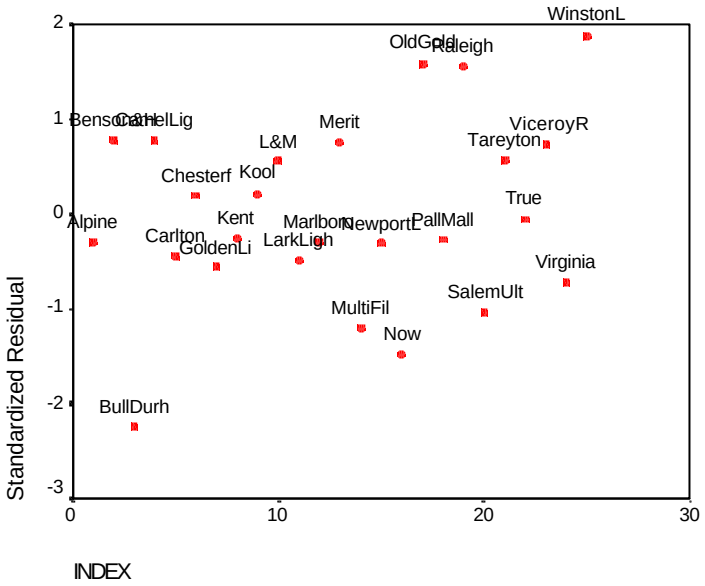
L'analisi può essere approfondita costruendo anche:

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 INDEX.

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

In questo caso, il grafico conferma ciò che avevamo osservato analizzando il diagramma di dispersione: l'osservazione 3 (Bull Durham), caratterizzata dal residuo più elevato in valore assoluto (pari a 2.235, dalla tabella **Casewise Diagnostics**), si discosta dalle altre (in questo senso è un *outlier*).



Casewise Diagnostics[a]

Case Number	NAME	Std. Residual	CO	Predicted Value	Residual
3	BullDurh	-2.235	23.50	26.5565	-3.0565

a. Dependent Variable: CO

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

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

Questo grafico può evidenziare un andamento nei residui che indica non linearità e può rivelare la presenza di punti outliers per la variabile esplicativa.

In questo caso il grafico non rivela alcun particolare andamento, a parte la presenza di un'osservazione della variabile esplicativa (sempre l'osservazione 3: Bull Durham) che è isolata rispetto al resto delle osservazioni.

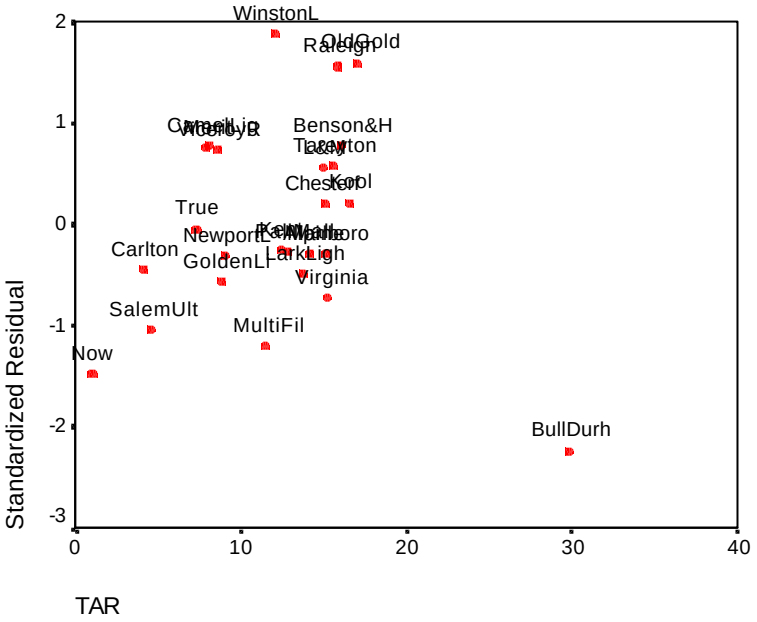
Appare evidente da questi grafici che l'osservazione Bull Durham è un *outlier* sia per la variabile dipendente, sia per la variabile indipendente.

A conferma di quanto osservato, per esercizio costruiamo un nuovo grafico (molto utile nel caso in cui l'identificazione di potenziali punti *outliers* (o **punti influenti**) sia più controversa rispetto al caso in esame) che ha
- sull'asse delle ascisse la statistica **Centered Leverage values**

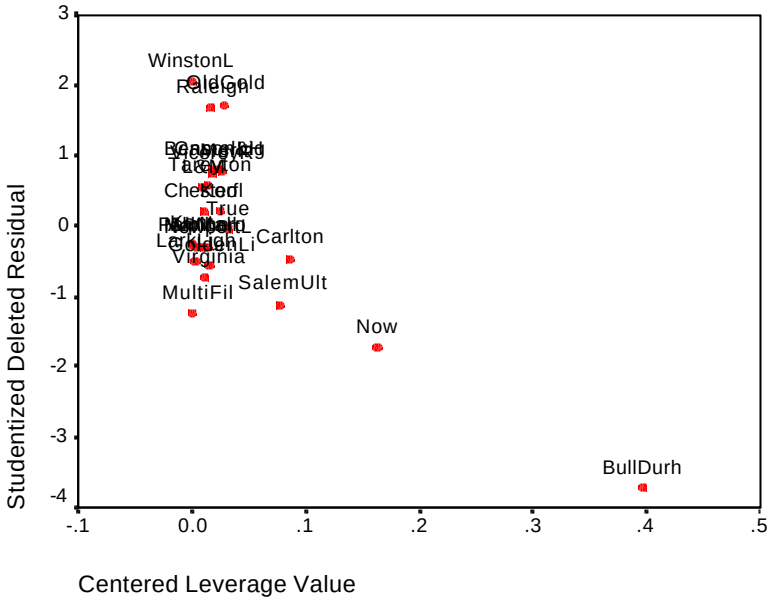
(*Il leverage di un caso è una misura della distanza tra il vettore contenente i valori delle variabili esplicative associate a quel caso e la media dei vettori contenenti i valori delle variabili esplicative associate a tutti i casi*)

- e sull'asse delle ordinate i **Residuals Studentized deleted** *(questi residui costituiscono uno strumento migliore per individuare potenziali outliers, in quanto i residui standardizzati tendono a sottovalutare la grandezza reale dei residui).*

Il grafico ottenuto nel nostro caso mostra chiaramente che l'osservazione Bull Durham è un *outlier* sia per la variabile dipendente, sia per la variabile indipendente (ha infatti leverage pari a 0.44233[1]), quindi si tratta di un potenziale **punto influente** (cioè un punto che può influire sulla stima dei coefficienti di regressione)



[1] 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.

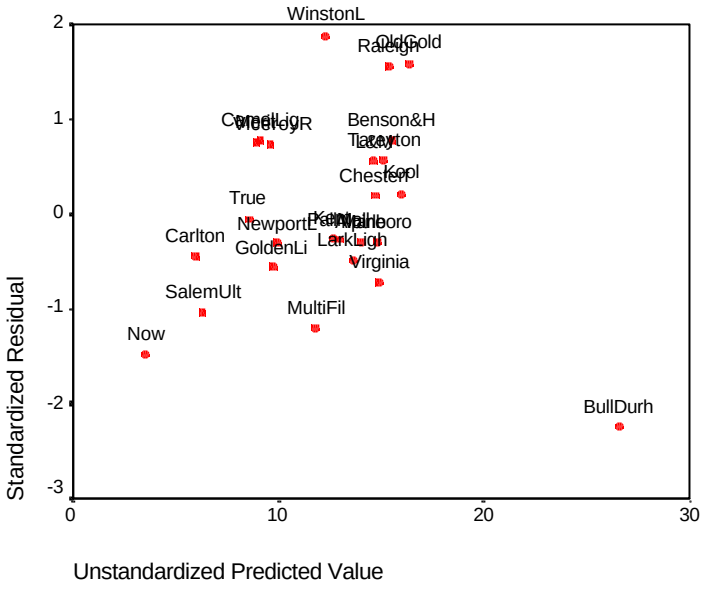


3. il plot dei residui standardizzati rispetto ai valori stimati.

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, a parte la presenza dell'outlier Bull Durham, isolato rispetto al resto dei dati.



- Come avevamo osservato subito guardando il diagramma di dispersione, e come l'analisi dei residui ha confermato, l'osservazione Bull Durham influenza in modo rilevante la stima dei coefficienti della retta interpolante. Per esercizio, studiamo la presenza di punti influenti anche mediante la distanza di Cook.

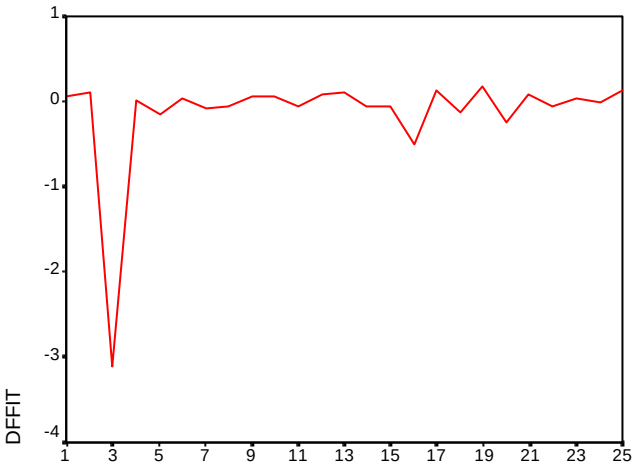
La distanza di Cook misura l'influenza di un singolo caso sulla stima dei coefficienti di regressione, quando il singolo caso viene rimosso dal processo di stima. Un valore della distanza di Cook >1 indica che il punto è influente.

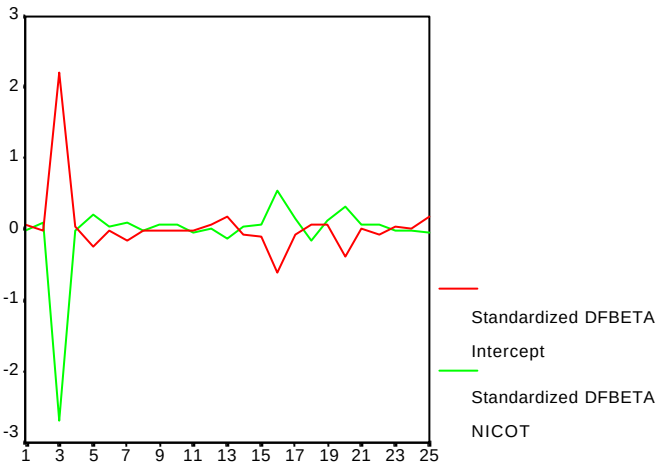
In questo caso il valore della distanza di Cook per l'osservazione Bull Durham è 2.98, mentre tutti gli altri casi hanno distanza di Cook <1.

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.

Per rappresentare graficamente i valori assunti da queste statistiche utilizziamo dal menù **Graphs** => **Sequence** .

Dai grafici emerge chiaramente che se l'osservazione Bull Durham influenza la stima di entrambi i parametri del modello.


DFFIT
1
0
-1
-2
-3
-4
1 3 5 7 9 11 13 15 17 19 21 23 25
Sequence number


3
2
1
0
-1
-2
-3
1 3 5 7 9 11 13 15 17 19 21 23 25
Sequence number
Standardized DFBETA Intercept
Standardized DFBETA NICOT

Domanda 2

Costruire un modello di regressione lineare multipla utilizzando come variabile dipendente CO e come variabili esplicative TAR e NICOT. Commentare i risultati ottenuti.

Analisi

Ipotizziamo che valga il modello

$$CO = b_0 + b_1 \cdot TAR + b_2 \cdot NICOT + e$$

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

- **Statistics** => **Descriptives**, **Collinearity diagnostics** 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, i variabili PRE_1 (che contiene i valori stimati) e la variabile ZRE_1 (che contiene i residui standardizzati).

Plots => attiviamo **Normal probability plot**

Analisi dell'output

La tabella **ANOVA** contiene la somma dei quadrati del modello di regressione (Regression), la somma dei quadrati dei residui (Residuals) e la somma dei quadrati totali (Total).
La **statistica F**, data da

$$F = \frac{\text{Regression}/(n-k-1)}{\text{Residuals}/(n-1)}$$

dove $k = 2$ (numero di variabili esplicative che compaiono nel modello), è altamente significativa (con un p-value prossimo a zero), perciò si rifiuta l'ipotesi nulla del test

$$H_0: \ b_1 = b_2 = 0 \quad \text{contro } H_1: \ b_i \neq 0 \ \text{ per almeno un } i$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	497.316	2	248.658	130.766	.000[a]
	Residual	41.834	22	1.902		
	Total	539.150	24			

a. Predictors: (Constant), NICOT, TAR

b. Dependent Variable: CO

La tabella **Coefficients** contiene le stime dei parametri del modello (B), gli errori standard degli stimatori ottenuti con il metodo dei minimi quadrati (Std.Error) e le statistiche (t) e i p-values (Sig.) dei test di Students che verificano se i parametri siano significativamente diversi da zero. Entrambi i parametri b_0 e b_1 risultano significativamente differenti da zero. Il test per valutare la significatività di b_2 porta invece ad accettare l'ipotesi nulla $H_0 : b_2 = 0$.

Multicollinearità.

Dal momento che la matrice di correlazione mette in evidenza che le variabili TAR e NICOT sono fortemente correlate (0.978), ci aspettiamo che vi siano problemi di **multicollinearità.**

Nella tabella **Coefficients** leggiamo i valori delle statistiche **Collinearity Statistics: Tolerance** e **VIF**.

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

$$Tolerance = 1 - R_i^2$$

dove R_i^2 è il coefficiente di correlazione multipla tra la variabile i-esima 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.

La statistica ***VIF (Variance Inflation Factor)*** *è il reciproco della statistica* ***Tolerance.*** *Un valore soglia per la statistica VIF è rappresentato da 10, che corrisponde a una Tolerance di 0.10)*

In questo caso i valori di **Tolerance** associati a TAR e NICOT sono al di sotto del valore di soglia, quindi siamo in presenza di multicollinearità.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.105	.821		3.782	.001		
	TAR	.974	.227	1.168	4.294	.000	.048	20.989
	NICOT	-2.867	3.642	-.214	-.787	.440	.048	20.989

a. Dependent Variable: CO

Al fine di verificare se si è in presenza di un problema di multicollinearità , SPSS calcola anche i condition indices e la matrice coefficient variance-decomposition.

Un condition index molto grande (maggiore di 30) indica un elevato grado di collinearità. La matrice coefficient variance-decomposition mostra la proporzione di varianza per ogni coefficiente di regressione (e quindi per ogni variabile esplicativa) attribuibile ad ogni condition index.

La tabella riportata sotto mostra che il *condition index* associato al modello con due regressori è molto vicino al valore di soglia e vi è una riga della matrice con le proporzioni che superano 0.90 per entrambi i coefficienti, per cui concludiamo che le variabili TAR e NICOT sono collineari.

Collinearity Diagnostics [a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	NICOT	TAR
1	1	2.891	1.000	.01	.00	.00
	2	.106	5.228	.71	.01	.02
	3	3.616E-03	28.275	.27	.99	.98

a. Dependent Variable: CO

Vi sono diversi "rimedi" al problema della multicollinerità, che tuttavia qui non approfondiremo. Un primo passo è chiedersi se sia migliore il modello con due repressori (collineari!), o un modello più semplice con un solo repressore.
Sappiamo che una misura della bontà del modello è data dal coefficiente di determinazione multipla R^2

$$R^2 = \frac{\text{Regression}}{\text{Total}} = 1 - \frac{\text{Residual}}{\text{Total}}$$

Tuttavia, R^2 cresce all'aumentare del numero di repressori; perciò, è preferibile considerare l'indice "R^2 aggiustato" (*adjusted* R^2)

$$R^2_{ad} = 1 - \frac{\text{Residual}/(n\text{-}k\text{-}1)}{\text{Total}/(n\text{-}1)} \text{ (Adjusted R square)}$$

che tiene conto del numero k di regressori.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.960[a]	.922	.915	1.3790

a. Predictors: (Constant), NICOT, TAR

b. Dependent Variable: CO

Per il modello lineare con il solo regressore TAR (si veda Domanda 1) si era trovato $R^2 = 0.92$ e $R^2_{ad} = 0.917$. Per questo modello con due repressori, NICOT e TAR, vediamo dalla tabella sopra riportata che l'indice R^2_{ad} risulta leggermente peggiore (0.915).
Non è sorprendente che il valore di R^2_{ad} rimanga elevato, nonostante la presenza di multicollinearità. La multicollinearità tuttavia rende molto instabili le stime dei coefficienti di correlazione e rende del tutto ambigua l'interpretazione del coefficiente di regressione come variazione della variabile dipendente in corrispondenza ad un incremento unitario della variabile esplicativa, quando le rimanenti variabili esplicative sono mantenute costanti.
Perciò concludiamo che è preferibile il modello con il solo regressore TAR.