Esercizio 2
(regressione lineare semplice con variabili qualitative)

DATI. Il data set gas.sav (fonte: http://www.statsci.org) contiene dati raccolti nel 1960 in una casa nel sud-est dell'Inghilterra. Sono stati misurati il consumo settimanale di gas (in 1000 cubic feet[1]) e la temperatura media esterna (in gradi Celsius) per 26 settimane prima di aver istallato un'intercapedine di isolamento nei muri e 18 settimane dopo tale istallazione. Le variabili sono

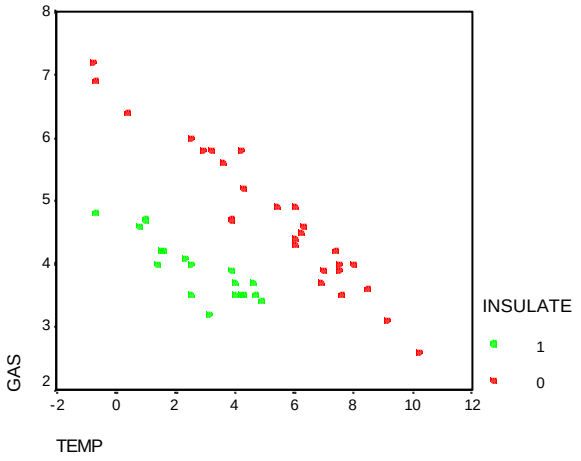
INSULATE	0=prima, 1=dopo
TEMP	Temperatura media esterna (°C)
GAS	Consumo di gas (in 1000 cubic feet)

Domanda

Studiare la dipendenza del consumo di gas (GAS) dalla temperatura (TEMP) tramite un modello di regressione lineare (software: 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 GAS, come **X-axis** la variabile TEMP, come **Set Markers by** INSULATE.



Dal diagramma di dispersione appare evidente che siamo in presenza di due gruppi di dati, relativi a prima e dopo l'installazione dell'intercapedine di isolamento. Per costruire un modello adeguato a rappresentare entrambi i gruppi di dati introduciamo come regressore la **variabile *dummy*** INSULATE (=0 prima dell'installazione dell'intercapedine di isolamento, =1 dopo).

Ipotizziamo dunque che valga il seguente modello:

$$GAS = b_0 + b_1 \cdot INSULATE + b_2 \cdot TEMP + e$$

e supponiamo che siano soddisfatte le ipotesi forti.

[1] Cubic foot : misura di volume corrispondente a 28.318 dm^3

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

- **Statistics** => **Descriptives** e dalla finestra **Residuals**, la voce **Casewise Diagnostics**, con **Outliers outside:** 1 **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** => **Histogram** e **Normal probability plot**

Analisi dell'output

La tabella **Descriptive Statistics** contiene media e deviazione standard delle variabili GAS e TEMP. Il consumo medio di gas è 4397.7 cubic feet , mentre la temperatura media esterna è 4.3114 °C.

Descriptive Statistics

	Mean	Std. Deviation	N
GAS	4.3977	1.0285	44
TEMP	4.3114	2.7260	44

In particolare, scegliendo dal menù **Analyze** => **Reports** =>**Reports summaries in rows:** come **Data columns** GAS, come **Break columns** INSULATE e da **Summary**, nella finestra **Break columns** => **Mean of values**, si ottiene l'output

INSULATE	GAS
0	
Mean	4.75
1	
Mean	3.89

Come ci si aspetta il consumo medio di gas diminuisce considerevolmente dopo l'installazione dell'intercapedine di isolamento.

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), le statistiche (t) e i p-values (Sig.) dei test di Students che verificano se i parametri siano significativamente diversi da zero.
Le statistiche dei test di Student assumono valori elevati in valore assoluto, perciò si può concludere che i parametri del modello sono tutti significativamente diversi da zero. Ciò risulta confermato dai corrispondenti p-values che sono tutti prossimi a zero.

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.717	.117		57.478	.000
	TEMP	-.368	.019	-.975	-19.465	.000
	INSULATE	-1.795	.104	-.868	-17.333	.000

a. Dependent Variable: GAS

Il modello lineare stimato è

$$GAS = 6.717 - 1.795 \cdot INSULATE - 0.368 \cdot TEMP$$

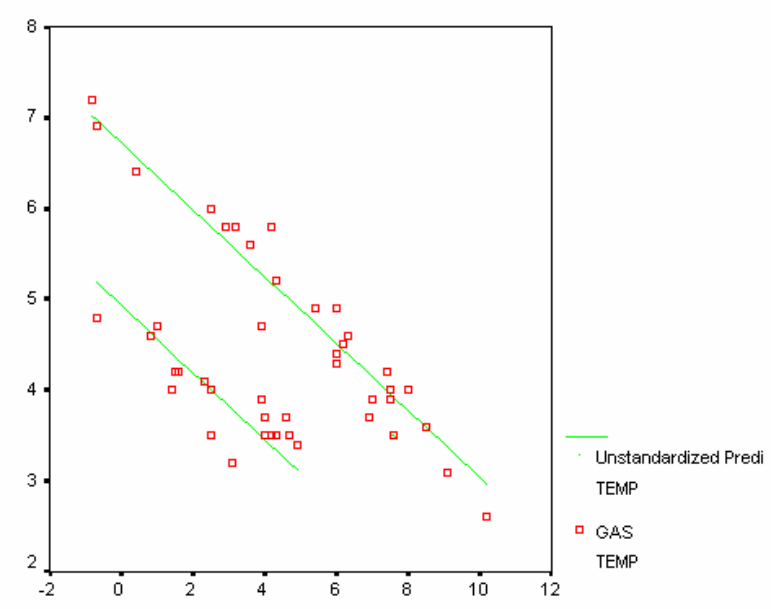
Poiché INSULATE assume i valori 0 oppure 1, il modello si può scrivere come

$$GAS = 6.717 - 0.368 \cdot TEMP \quad \text{se } INSULATE = 0 \quad \text{cioè prima dell'installazione}$$

$$GAS = 4.922 - 0.368 \cdot TEMP \quad \text{se } INSULATE = 1 \quad \text{cioè dopo l'installazione}$$

Il consumo di gas in corrispondenza a una temperatura di 0°C diminuisce considerevolmente dopo l'installazione dell'intercapedine (da 6717 cubic feet a 4922 cubic feet). Inoltre per ogni incremento di un grado Celsius della temperatura, il consumo di gas diminuisce di 368 cubic feet alla settimana.

Rappresentiamo ora sullo stesso grafico i valori osservati di GAS e TEMP e le rette interpolanti relative ai due modelli descritti sopra. Dal menu **Graphs** selezioniamo **Scatter** e quindi **Overlay** . Come prima coppia **Y-X Pairs** scegliamo le variabili GAS e TEMP, come seconda coppia **Y-X Pairs** le variabili PRE_1 e TEMP. Quindi clicchiamo sul diagramma di dispersione e dal menù scegliamo **Format** e quindi **Interpolation**



La capacità esplicativa della variabile TEMP di rappresentare la variabile dipendente GAS per mezzo di una retta (con coefficiente angolare e intercetta diversi, in corrispondenza ai dati prima e dopo l'installazione) 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 del coefficiente di determinazione R^2 che è pari a 0.9119. Quindi il modello spiega il 91.9% della variabilità della variabile dipendente GAS.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959[a]	.919	.915	.2993

a. Predictors: (Constant), INSULATE, TEMP

b. Dependent Variable: GAS

Il precedente modello presuppone che non vi sia effetto del carattere INSULATE sul coefficiente angolare. Se invece pensiamo che INSULATE possa avere effetto anche sulla sensibilità del consumo di gas alla temperatura, dobbiamo introdurre un'ulteriore variabile, come ora si mostrerà.
Allo scopo di valutare l'effetto del carattere INSULATE sul coefficiente angolare stimiamo il modello

$$GAS = b_0 + b_1 \cdot INSULATE + b_2 \cdot TEMP + b_3 \cdot IT + e$$

dove $IT = INSULATE \cdot TEMP$
Si ottiene il seguente output

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,967[a]	,936	,931	,2699

a. Predictors: (Constant), IT, TEMP, INSULATE

b. Dependent Variable: GAS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42,575	3	14,192	194,770	,000[a]
	Residual	2,915	40	7,286E-02		
	Total	45,490	43			

a. Predictors: (Constant), IT, TEMP, INSULATE

b. Dependent Variable: GAS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,854	,114		60,320	,000
	INSULATE	-2,263	,173	-1,094	-13,099	,000
	TEMP	-,393	,019	-1,042	-20,925	,000
	IT	,144	,045	,242	3,224	,003

a. Dependent Variable: GAS

Il modello lineare stimato è

$$GAS = 6.854 - 2.263 \cdot INSULATE - 0.393 \cdot TEMP + 0.144 \cdot IT$$

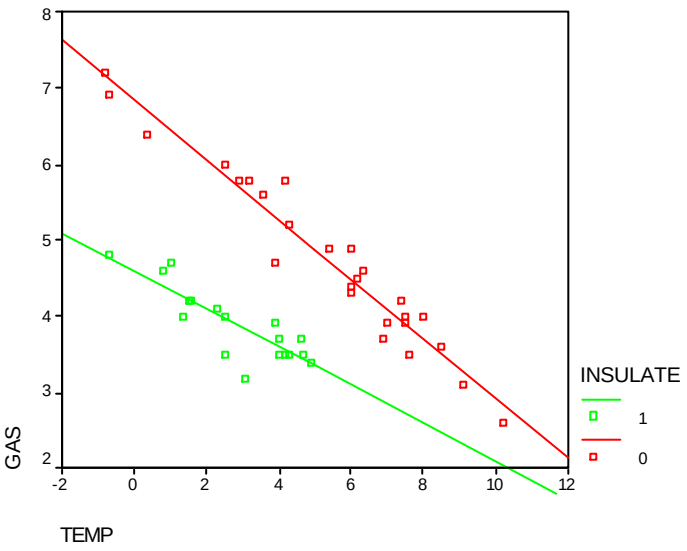
Poiché INSULATE assume i valori 0 oppure 1, il modello si può scrivere come

$GAS = 6.854 - 0.393 \cdot TEMP$ se $INSULATE = 0$ cioè prima dell'installazione

$GAS = 4.591 - 0.249 \cdot TEMP$ se $INSULATE = 1$ cioè dopo l'installazione

Si noti che ora le due rette hanno diversa intercetta e *anche diverso coefficiente angolare*.

Rappresentiamo ora sullo stesso grafico i valori osservati di GAS e TEMP e le rette interpolanti relative ai due modelli descritti sopra. Costruiamo il diagramma di dispersione => Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple** . Come **Y Axis** scegliamo la variabilee GAS, come **X Axis** la variabile TEMP e come **Set Markers by** la variabile INSULATE. Quindi clicchiamo sul grafico e dal menù scegliamo **Chart** => **Options**=>**Fit line Subgroups**



<u>Analisi dei residui.</u>

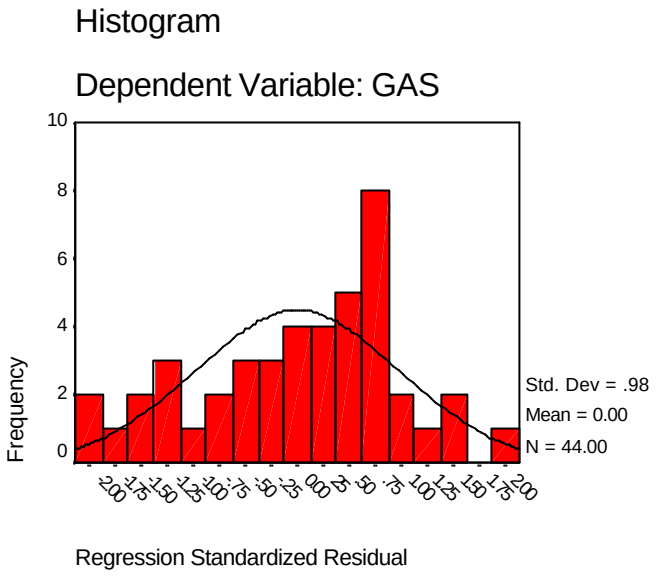
Un ulteriore strumento per controllare la bontà di un modello di regressione è dato dall'analisi dei residui. La svolgiamo qui per il primo dei due modelli stimati al punto precedente.

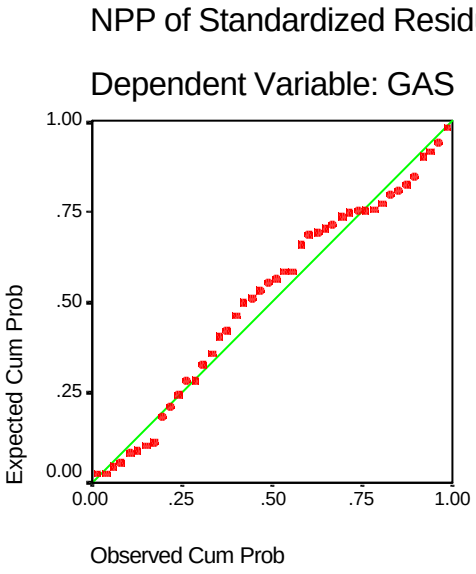
Se sono verificate le ipotesi forti

- 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 i residui seguono approssimativamente una distribuzione normale, sebbene sia riscontrabile una certa asimmetria nei dati. Il NPP dei residui standardizzati mostra

che i punti rappresentati 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

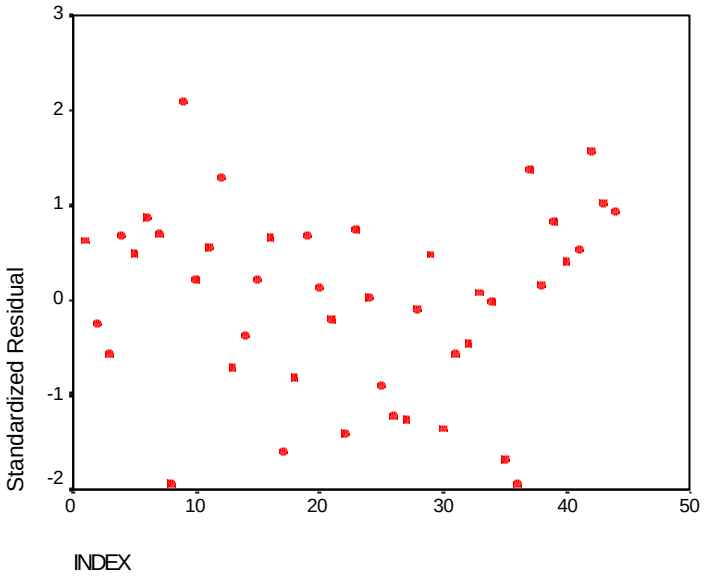
Costruiamo quindi

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 INDEX.

I residui standardizzati sono tutti compresi tra –2 e 3. In particolare vi sono 3 osservazioni con residui standardizzati prossimi a 2 in valore assoluto.

La tabella **Casewise Diagnostics** in cui compaiono i residui standardizzati che sono in valore assoluto maggiori di uno, individua le osservazioni 8 (-1.949),9 (2.096),36 (-1.947) come potenziali outliers



Casewise Diagnostics[a]

Case Number	Std. Residual	GAS	Predicted Value	Residual
8	-1.949	4.70	5.2831	-.5831
9	2.096	5.80	5.1728	.6272
12	1.300	4.90	4.5110	.3890
17	-1.604	3.70	4.1801	-.4801
22	-1.413	3.50	3.9227	-.4227
26	-1.225	2.60	2.9667	-.3667
27	-1.269	4.80	5.1799	-.3799
30	-1.362	4.00	4.4077	-.4077
35	-1.682	3.50	4.0033	-.5033
36	-1.947	3.20	3.7827	-.5827
37	1.375	3.90	3.4885	.4115
42	1.567	3.70	3.2311	.4689
43	1.021	3.50	3.1944	.3056

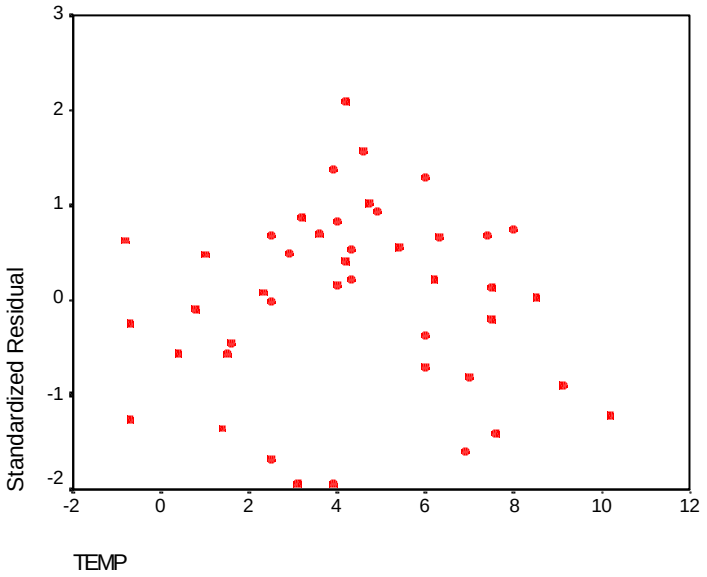
a. Dependent Variable: GAS

2. il plot dei residui standardizzati rispetto alle variabili esplicative TEMP e INSULATE

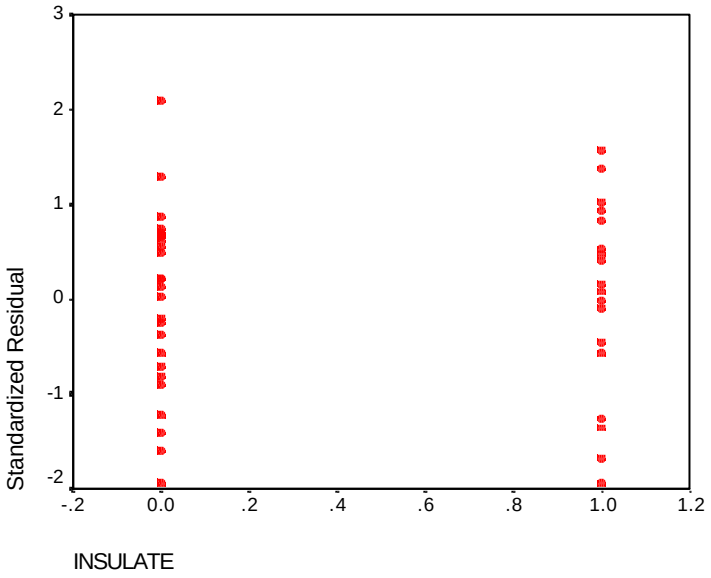
Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple**. Come **Y-axis** la variabile ZRE_1 e come **X-axis** la variabile TEMP.

Questo grafico può mostrare un andamento nei residui che indica non linearità.

In questo caso non è riconoscibile un comportamento specifico, anche se in corrispondenza a temperature vicine alla media della variabile TEMP, si hanno residui positivi, per cui il modello tende a sottostimare le osservazioni.



Dal menu **Graphs** selezioniamo **Scatter** e quindi **Simple**. Come **Y-axis** la variabile ZRE_1 e come **X-axis** la variabile INSULATE.
Questo grafico mostra che all'interno di ogni gruppo i residui si distribuiscono in modo abbastanza uniforme tra –2 e 2, con una prevalenza, per entrambi i gruppi, di residui positivi, per cui concludiamo, anche in questo caso, che il modello tende a sottostimare le osservazioni.



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.