Statistica descrittiva: analisi di regressione

- L'*analisi di regressione* permette di esplorare le relazioni tra due insiemi di valori (p.e. i valori di due attributi di un campione) alla ricerca di associazioni.

- Per esempio possiamo usare l'analisi di regressione per determinare se:
  - le spese in pubblicità sono associate con le vendite
  - il fumo è associato con le malattie cardiache
  - la dieta mediterranea è associata con la durata della vita

Scatter plots (diagrammi a punti)

- Un primo approccio all'analisi di regressione è la creazione di uno scatter plot, che mostra su un piano XY un punto per ogni coppia di valori

- Per esempio se abbiamo un campione che riporta per ciascuna famiglia le entrate mensili, le spese per attività culturali, le spese per attività sportive ecc., possiamo creare uno scatter plot che usa le coppie entrate-spese culturali per indagare l'esistenza di una relazione

Scatter plots in Excel

- Excel consente la creazione di scatter plots mediante lo strumento *chart wizard*
- Il chart wizard oltre a consentire la creazione del grafico a partire dalla selezione delle liste di valori di cui si vuole studiare l'associazione, consente:
  - di generare la *trendline*, ovvero la curva che meglio approssima l'andamento dell'insieme di valori sulle ordinate rispetto all'insieme di valori sulle ascisse
  - di generare l'*equazione di regressione* ovvero l'equazione della trendline
  - trendline e equazione di regressione ci permettono di classificare l'associazione: lineare, logaritmica, esponenziale ecc.

Esempio: entrate e spese familiari

- A partire dal file EXPENSES.XLS trovare:
  - l'associazione tra entrate e spese per cultura
  - l'associazione tra entrate e spese per sport
  - l'associazione tra spese per sport e spese per cultura

Misure di associazione: covarianza

- La *covarianza* quantifica la la forza della relazione tra due insiemi di valori, ovvero misura quanto lineare e` la dipendenza tra i due insiemi;
- La covarianza e` la media del prodotto delle deviazioni dei valori dalla media degli insiemi dei dati
- In formula:

$$\mathrm{cov}(X,Y) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (X_{\mathrm{i}} - \overline{X}) \times (Y_{\mathrm{i}} - \overline{Y})}{n}$$

- un valore positivo indica una variazione di X e Y nella stessa direzione, un valore negativo l'opposto

Misure di associazione: correlazione

- un limite della misura di covarianza come misura descrittiva e` la sua dipendenza dall'unita` di misura usata per i valori;
- per esempio possiamo gonfiare il fattore covarianza per un fattore 1000, semplicemente sostituendo come unita` di misura euro in luogo di migliaia di euro (naturalmente se le unita` sono appropriate)
- La misura di *correlazione* risolve il problema producendo un risultato indipendente dalle unita` di misura e compreso tra –1 e 1
- In formula

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Stdev}(X) \times \mathrm{Stdev}(Y)}$$

Misure di associazione: correlazione

- Un valore della correlazione e` vicino a –1 indica che i due insiemi di valori tendono a variare in senso opposto
- Un valore della correlazione vicino a +1 indica che i due insiemi di valori tendono a variare nello stesso senso
- Una indipendenza nelle variazioni dei due valori produce un indice di correlazione uguale a 0
- Ma, attenzione: l'indice di correlazione e` rilevante solo per relazioni *lineari*
- L'indice puo` risultare vicino a 0 anche se esiste una relazione non lineare tra i due insiemi di valori.

Coefficiente di correlazione (Pearson)

- $r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$

R^2 coefficiente di determinazione

- Misura la percentuale di variazione della variabile dipendente spiegata dalla variazione della variabile indipendente.
- Il range è 0..1
- Per esempio un valore 0,8 può essere interpretato come l'80% delle variazioni è spiegato dalle variazioni della variabile indipendente, il 20% possono esser dovute da variabilità random
- Nel caso di regressione lineare coincide con il quadrato del coefficiente di correlazione

Calcolo del coefficiente di determinazione

- $R^2 = {}^{ESS}/_{TSS} = 1 - {}^{RSS}/_{TSS}$

$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ devianza spiegata dal modello

$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2$ devianza totale

$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ devianza residua

Con y_i dati osservati

$\overline{y}$ media dei dati osservati

$\hat{y}_i$ dati previsti dal modello

Regressione lineare

- Se i dati di uno scatter plot cadono approssimativamente su una retta, la *regressione lineare* consente di calcolare la migliore retta che approssima i dati
- La retta di regressione è descritta da una equazione

  $y = a + bx$

  dove y è la variabile dipendente e x la variabile indipendente

  a e b i coefficienti, rispettivamente il termine costante e b il coefficiente angolare (slope)
- I residui sono le differenze tra i valori dati e quelli stimati dalla retta

Calcolo dei coefficienti

- L’idea è di trovare una retta che minimizzi la somma dei residui al quadrato ovvero la distanza totale dei valori osservati dai valori stimati, al quadrato per evitare il condizionamento dsel segno.

Somma dei residui al quadrato

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Metodo dei minimi quadrati

- $b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

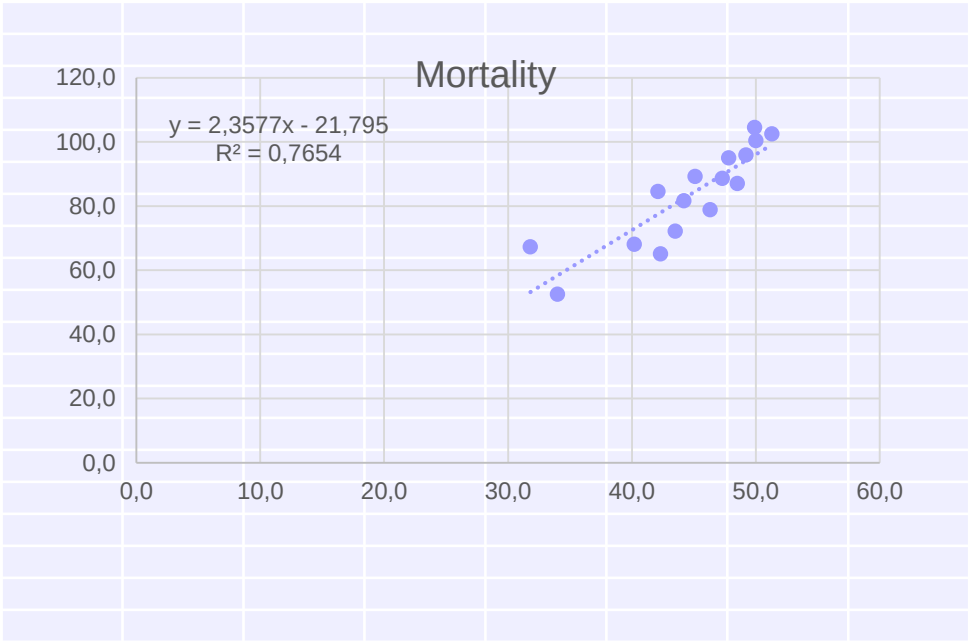
- $a = \overline{y} - b\overline{x}$

Statistiche di regressione

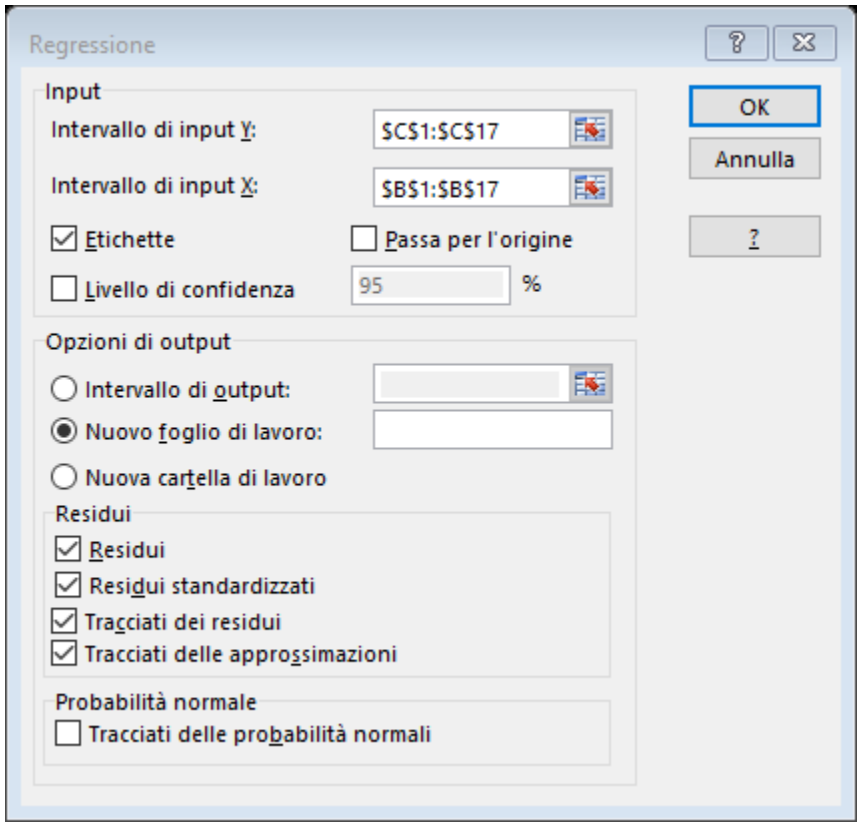
- Statistiche finalizzate a valutare l'adeguatezza di un modello lineare
- E' possibile usare Analysis Toolpack di Excel per calcolare le statistiche di regressione.
- Si faccia riferimento al file BCANCER:
  - Contiene dati di uno studio del 1965 che analizza le relazioni tra la temperatura media annuale e la percentuale di mortalità per certi tipi di cancro al seno.
  - I soggetti dell'analisi provengono da 16 regioni diverse in Gran Bretagna, Norvegia e Svezia

Region	Temperature	Mortality	
1	31,8	67,3	
2	34,0	52,5	
3	40,2	68,1	
4	42,1	84,6	
5	42,3	65,1	

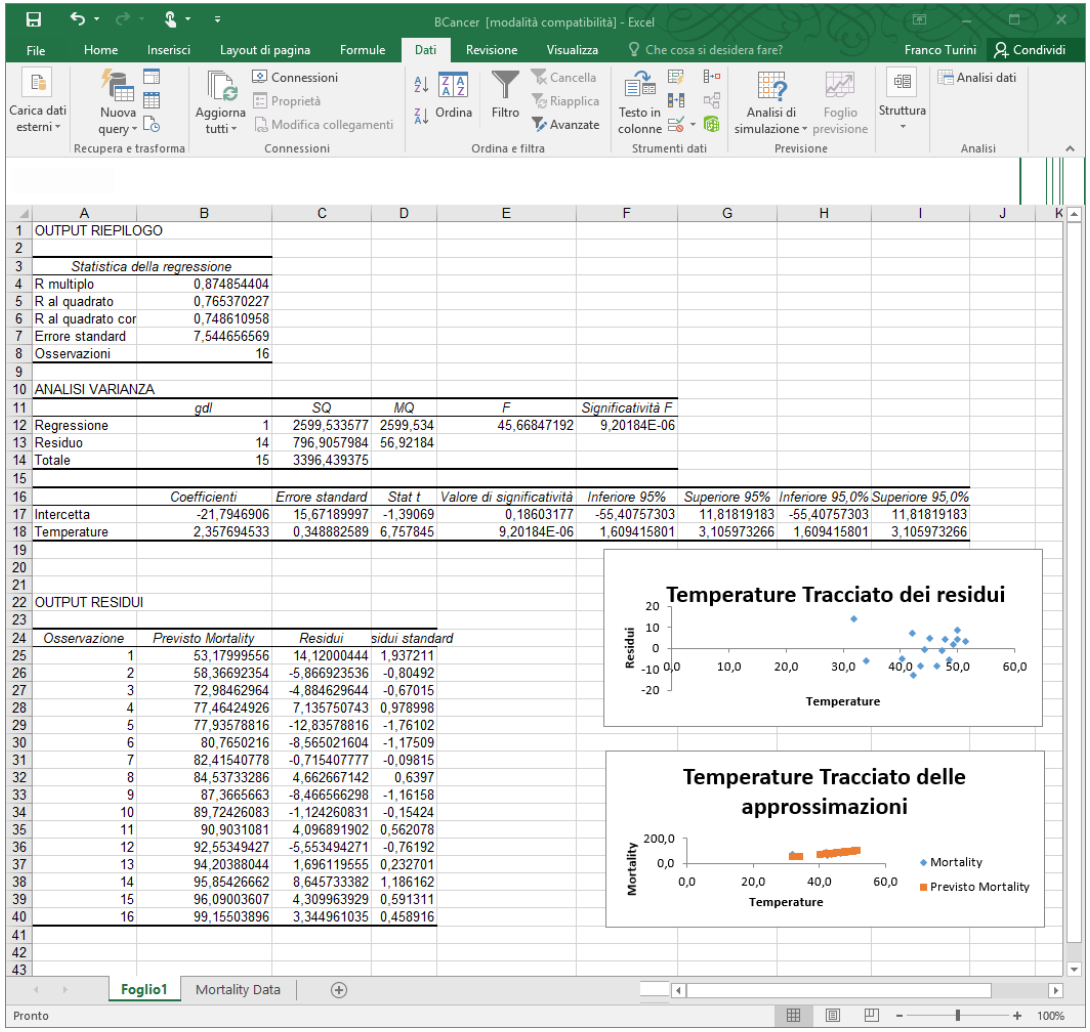
Calcolo della retta di regressione



Statistiche della regressione usando la funzione Regression dell'Analysis Toolpack



Output dell'analisi di regressione



BCancer [modalità compatibilità] - Excel

	A	B	C	D	E	F	G	H	I	J	K
1	OUTPUT RIEPILOGO										
2											
3	*Statistica della regressione*										
4	R multiplo	0,874854404									
5	R al quadrato	0,765370227									
6	R al quadrato cor	0,748610958									
7	Errore standard	7,544656569									
8	Osservazioni	16									
9											
10	ANALISI VARIANZA										
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*					
12	Regressione	1	2599,533577	2599,534	45,66847192	9,20184E-06					
13	Residuo	14	796,9057984	56,92184							
14	Totale	15	3396,439375								
15											
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*	*Inferiore 95,0%*	*Superiore 95,0%*		
17	Intercetta	-21,7946906	15,67189997	-1,39069	0,18603177	-55,40757303	11,81819183	-55,40757303	11,81819183		
18	Temperature	2,357694533	0,348882589	6,757845	9,20184E-06	1,609415801	3,105973266	1,609415801	3,105973266		
19											
20											
21											
22	OUTPUT RESIDUI										
23											
24	*Osservazione*	*Previsto Mortality*	*Residui*	*sidui standard*							
25	1	53,17999556	14,12000444	1,937211							
26	2	58,36692354	-5,866923536	-0,80492							
27	3	72,98462964	-4,884629644	-0,67015							
28	4	77,46424926	7,135750743	0,978998							
29	5	77,93578816	-12,83578816	-1,76102							
30	6	80,7650216	-8,565021604	-1,17509							
31	7	82,41540778	-0,715407777	-0,09815							
32	8	84,53733286	4,662667142	0,6397							
33	9	87,3665663	-8,466566298	-1,16158							
34	10	89,72426083	-1,124260831	-0,15424							
35	11	90,9031081	4,096891902	0,562078							
36	12	92,55349427	-5,553494271	-0,76192							
37	13	94,20388044	1,696119555	0,232701							
38	14	95,85426662	8,645733382	1,186162							
39	15	96,09003607	4,309963929	0,591311							
40	16	99,15503896	3,344961035	0,458916							
41											
42											
43											

Foglio1 | Mortality Data

Pronto

100%

Output dell’analisi di regressione

- L’output è organizzato su 5 aree:
  - Statistiche di regressione
  - Analisi di varianza (ANOVA)
  - Stima dei parametri
  - Residui
  - grafici

Statistiche di regressione

Statistica della regressione	
R multiplo	0,874854404
R al quadrato	0,765370227
R al quadrato corretto	0,748610958
Errore standard	7,544656569
Osservazioni	16

- R-multiplo è la radice quadrata di R al quadrato, ed è uguale al valore assoluto della correlazione tra la variabile dipendente e la variabile predittore
- R al quadrato corretto viene calcolato in caso di regressione con più di un predittore
- L'errore standard misura la tipica deviazione di un valore osservato (x,y) dalla retta di regressione (media delle deviazioni dalla retta di regressione).
- La formula dell'errore standard per un campione è

$$\sqrt{\frac{\sum(y - y')^2}{n-2}}$$

- Dove n è la numerosità del campione, y è il valore osservato e y' il valore atteso.

Analisi di varianza

ANALISI VARIANZA					
	gdl	SQ	MQ	F	Significatività F
Regressione	1	2599,533577	2599,534	45,66847192	9,20184E-06
Residuo	14	796,9057984	56,92184		
Totale	15	3396,439375			

- La tabella di Analisi di Varianza (ANOVA) analizza la variabilità dell'indice di mortalità La variabilità è divisa in due parti: la prima è la variabilità dovuta alla retta di regressione e la seconda dovuta a variabilità casuale (random)
- Gdl indica il grado di libertà per ciascuna parte. La somma dei due è uguale al numero di osservazioni meno 1. Nell'esempio un grado di libertà è attribuito alla retta di regressione e 14 alla variabilità random.
- SQ è la somma dei quadrati, ovvero la somma delle deviazioni al quadrato della variabile dipendente dalla media divisa in due parti: somma dei quadrati di regressione (somma delle deviazioni quadrate tra la retta di regressione e la media) e somma dei quadrati residua (somma delle deviazioni quadrate della variabile dipendente dalla retta di regressione)

Analisi di varianza (Cont.)

ANALISI VARIANZA					
	gdl	SQ	MQ	F	Significatività F
Regressione	1	2599,533577	2599,534	45,66847192	9,20184E-06
Residuo	14	796,9057984	56,92184		
Totale	15	3396,439375			

- La percentuale della somma dei quadrati totale che viene attribuita alla regressione è il valore R quadro, che misura la percentuale di variabilità spiegato dalla retta di regressione
- MQ è la media dei quadrati ovvero la somma dei quadrati divisa per i gradi di libertà ed è uguale al quadrato dell'errore standard
- F (F-ratio) è il rapporto del quadrato medio della regressione (2599,5) con l'errore quadratico medio dei residui (56,9). Un grande valore di F indica che la regressione può essere statisticamente significativa

Stima dei parametri

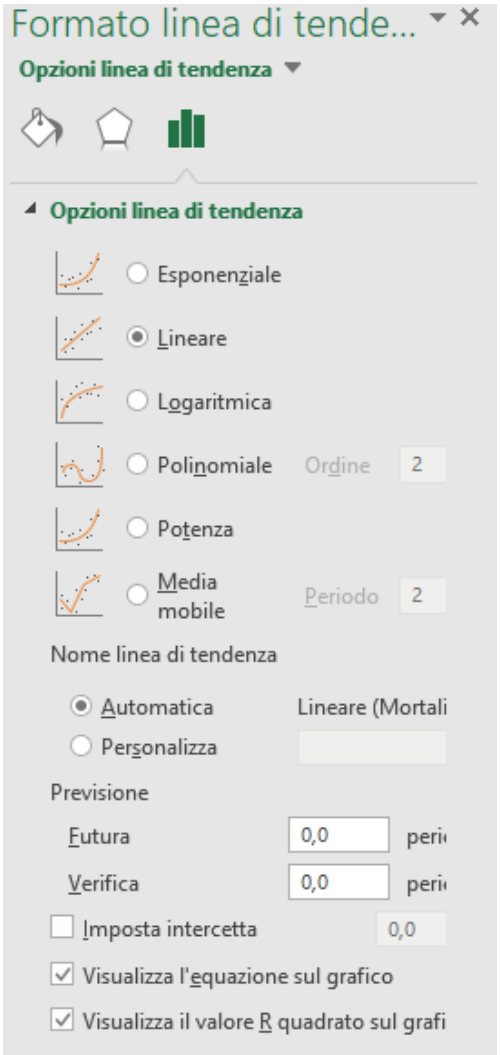
	Coefficienti	Errore standard	Stat t	Valore di significatività	Inferiore 95%	Superiore 95%	Inferiore 95,0%	Superiore 95,0%
Intercetta	-21,7946906	15,67189997	-1,39069	0,18603177	-55,40757303	11,81819183	-55,40757303	11,81819183
Temperature	2,357694533	0,348882589	6,757845	9,20184E-06	1,609415801	3,105973266	1,609415801	3,105973266

- La prima colonna riporta i valori calcolati del termine noto (intercetta) e del coefficiente angolare (Temperatura nell'esempio)
- La colonna successiva riporta gli errori standard per questi due valori stimati in accordo a una distribuzione t con n-2 gradi di libertà (vedi la parte di statistica inferenziale)
- I valori significativi sono gli intervalli stimati al 95% di confidenza per il termine noto [-55,4;11,8] e [1,6;3,1] per il coefficiente angolare.
- Conclusione per l'esempio dato:

  possiamo affermare con confidenza del 95% che per ciascun grado di aumento della temperatura annuale l'indice di mortalità della regione aumenta di un valore nell' intervallo [1,6;3,1]

La retta come modello

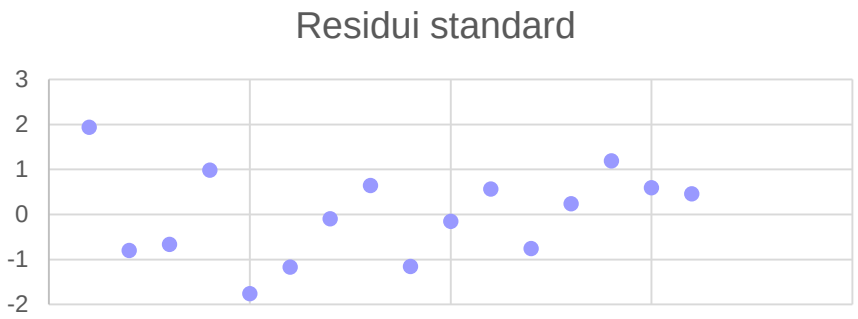
- Fare il fitting con altre possibili curve della trendline



E verificare se altre hanno un R-quadro migliore

Distribuzione normale dei residui

- I residui normalizzati dovrebbero distribuirsi in accordo a una normale con media 0 e deviazione standard 1
- I residui normalizzati dovrebbero cadere nell'intervallo [-2; 2] (Nota: il valore preciso dipende dalla t-distribution e dai gradi di libertà).



- Come si vede un solo residuo è ai bordi dell'intervallo di accettazione.
- Non ci sono outlier, ovvero valori con un residuo al di fuori dell'intervallo di $[-2\sigma;2\sigma]$ che in accordo alla distribuzione normale contiene il 95% dei valori.

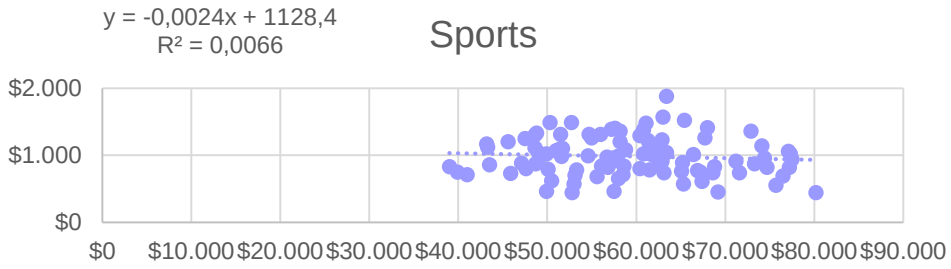
Ancora un esempio

- Vedi EXPENSES

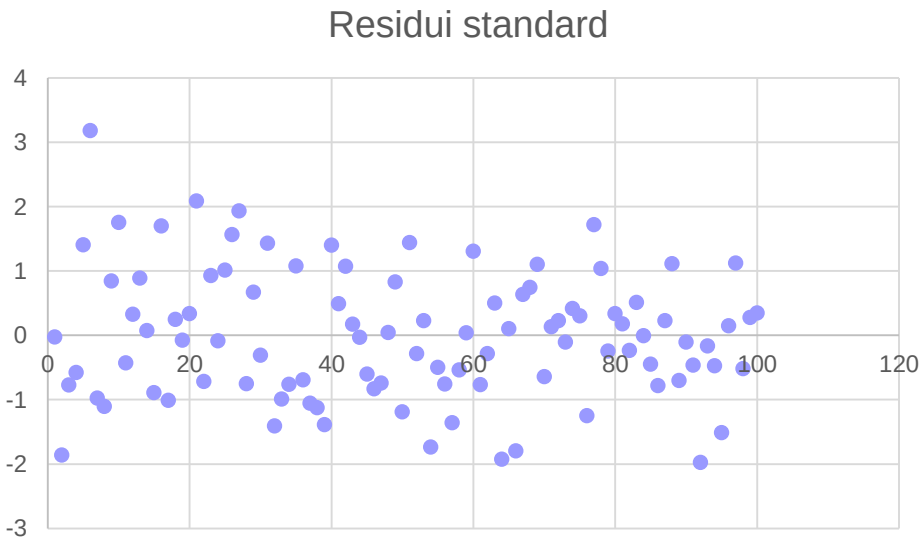
Salary	Culture	Sports	Dining
$54.600	$1.020	$990	$1.510
$57.500	$1.100	$460	$1.180
$53.300	$900	$780	$1.590
$43.500	$570	$860	$1.750
$57.200	$900	$1.390	$2.120
$63.400	$820	$1.880	$3.090
$58.500	$1.340	$710	$1.540
$55.600	$1.250	$680	$1.800
$61.300	$1.190	$1.220	$2.330
$61.100	$640	$1.480	$2.670
$77.200	$900	$820	$2.850
$58.800	$710	$1.080	$2.200
$62.900	$1.240	$1.230	$2.430
$61.900	$1.270	$1.000	$2.110
$76.500	$1.180	$690	$1.820
$50.300	$810	$1.490	$2.100
$45.900	$840	$730	$920

- Studiamo la regressione tra Salary e Sports

Diagramma XY e retta di regressione



Plot dei residui normalizzati



Limiti del coefficiente di correlazione

- il coefficiente di correlazione è suscettibile alla presenza di outlier (verificare con esempi che l'eliminazione di outliers incrementa il coefficiente di correlazione di Pearson
- Il calcolo del coefficiente assume una relazione lineare

Correlation is not necessarily causation

Una forte correlazione con buone statistiche per la regressione non è una prova di causalità, che può solo essere dimostrata da ragionamenti legati alla natura del fenomeno da studiare.

Se no a cosa servirebbero sociologi, economisti e scienziati in genere?

Esercitazione (1)

- Il responsabile del personale della Beta Tecnologies Inc. sta cercando di individuare la variabile che meglio spiega le variazioni di stipendio degli impiegati usando un campione che riporta i dati di 52 impiegati a tempo pieno.
- I dati sono nel file IMPIEGATI.XLS.
- Si generino diagrammi XY per determinare quale delle seguenti variabili ha la relazione lineare *più forte* con lo stipendio annuale:
  - sesso
  - età
  - numero di anni di esperienza lavorativa prima dell'assunzione in azienda
  - numero di anni di impiego in azienda
  - numero di anni di educazione post-secondaria.