

Spese Mediche USA

Spese mediche coperte dal piano assicurativo negli USA

A. Cola, M. Simonetti, R. Urso

22 maggio 2024

- 1 Introduzione
- 2 Analisi esplorativa
- 3 Modello di regressione
- 4 Verifica modello
- 5 Conclusioni

Introduzione

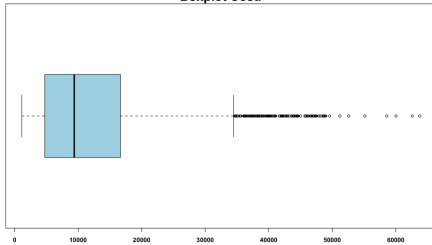
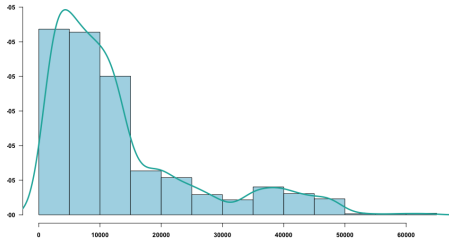
- Il dataset¹ preso in esame riguarda le spese mediche annuali sostenute negli Stati Uniti d'America da individui iscritti ad un piano assicurativo.
- È costituito da 1338 osservazioni sulle quali sono state rilevate 7 variabili di cui 4 quantitative e 3 qualitative: età, sesso, BMI, figli, fumatore, regione, costi.
- L'obiettivo ultimo è quello di individuare quali sono le variabili che spiegano in misura maggiore le spese mediche degli individui.

¹<https://www.kaggle.com/datasets/mirichoi0218/insurance> 

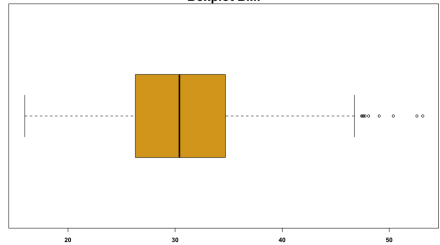
Considerazioni iniziali

- Prima di procedere con la stima del modello, è stato effettuato un lavoro di pulizia del dataset dal quale sono stati rimossi un duplicato e diversi *outlier* individuati nelle unità oltre il baffo superiore del boxplot relativo al BMI.
- L'eliminazione degli outlier della variabile costi comporterebbe la rimozione dal dataset di quello che sembra essere il gruppo dei **fumatori obesi**, come si può osservare nello *scatter plot* delle prossime slide.
- Per tale motivo si è deciso di non rimuovere ulteriori unità statistiche che potrebbero essere rappresentative di uno specifico sottogruppo della popolazione.

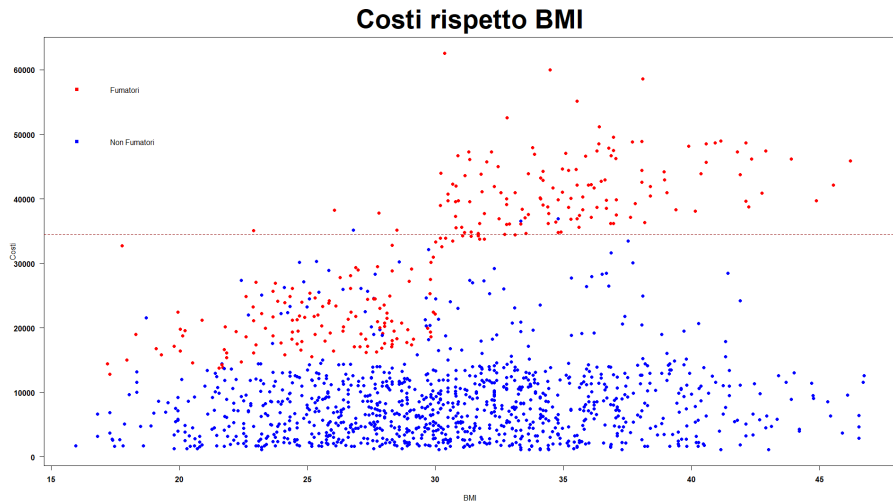
Boxplot Costi

*Distribuzione dei costi*

Boxplot BMI

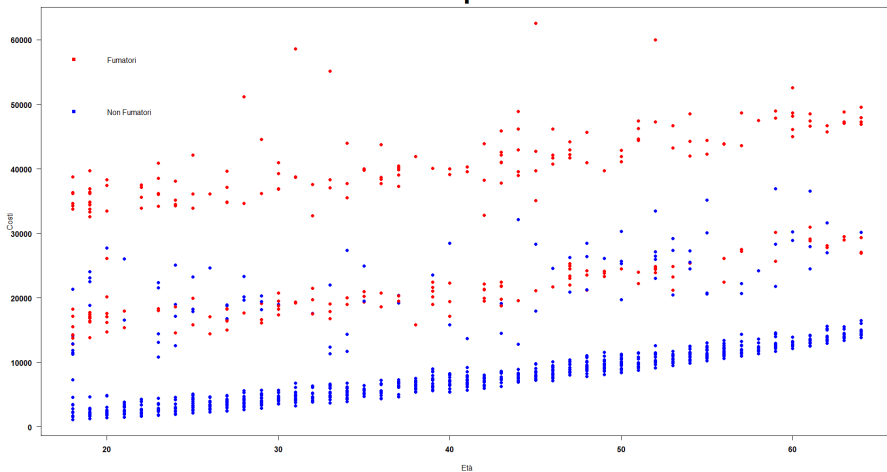
*Distribuzione del bmi*

Relazione tra costi e BMI

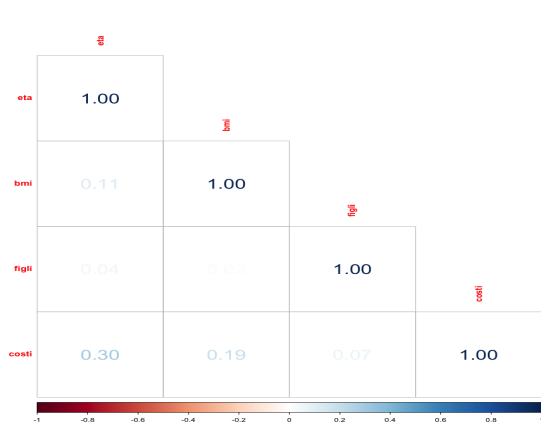


Relazione tra costi ed età

Costi rispetto Età



Matrice di correlazione e associazione



	fumo	regione	Sesso
fumo	1.000	0.026	0.080
regione	0.026	1.000	0.009
Sesso	0.080	0.009	1.000

Trasformazione dei costi

- Dall'analisi esplorativa emerge una forte asimmetria della distribuzione dei costi.
- Per ridurre tale asimmetria è stata utilizzata la classe di trasformazioni di Box-Cox, definita come:

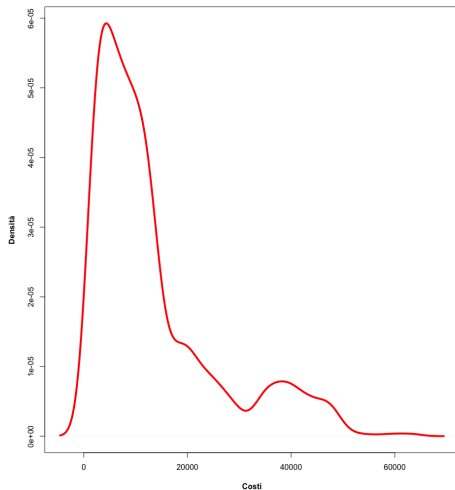
$$Z_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda}, & \text{se } \lambda \neq 0; \\ \log(Y_i), & \text{se } \lambda = 0; \end{cases} \quad i = 1, 2, \dots, 1329; \quad (1)$$

con Y_i variabile d'interesse.

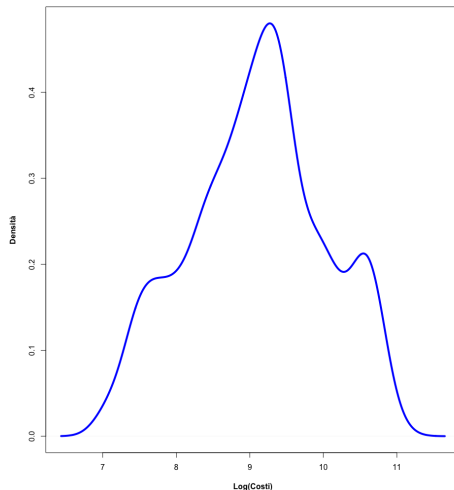
- Avendo stimato $\lambda \simeq 0$, è stato applicato il logaritmo alla variabile costi.

Confronto distribuzioni

Istogramma perequato costi



Istogramma perequato log(costi)



Specificazione modello

- Per spiegare i costi in funzione delle variabili restanti, verrà stimato il seguente modello di regressione lineare multipla:

$$Y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_6 x_{i6} + \epsilon_i, \quad i = 1, 2, \dots, 1329; \quad (2)$$

avendo posto pari a Y_i^* la trasformazione logaritmica di Y_i .

- Per la stima del modello la variabile regione è stata dicotomizzata, il che renderà pari a quattro il numero di dummy presenti nella specificazione.

Stima modello

Di seguito, l'output prodotto da R in seguito alla stima del modello:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.8091014	0.0766056	88.885	< 2e-16	***
eta	0.0345318	0.0008749	39.468	< 2e-16	***
sesto	0.0730581	0.0244761	2.985	0.00289	**
bmi	0.0136662	0.0021312	6.412	1.99e-10	***
figli	0.1009646	0.0101024	9.994	< 2e-16	***
fumo	1.5475704	0.0303563	50.980	< 2e-16	***
regione	0.1114527	0.0250500	4.449	9.34e-06	***

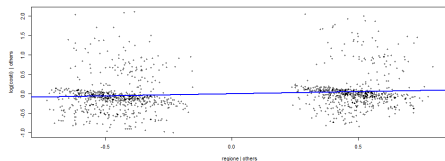
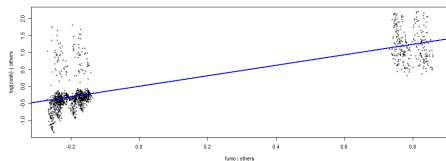
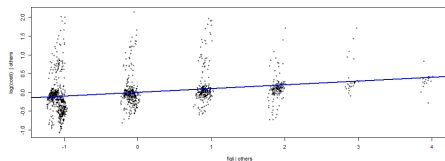
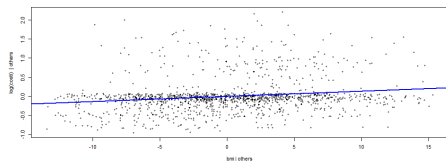
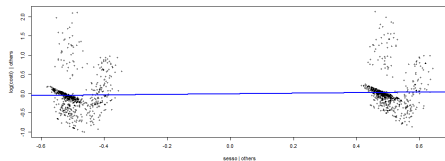
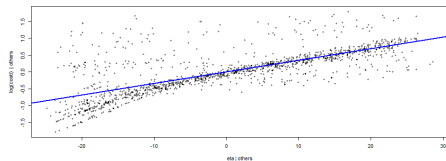
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4442 on 1322 degrees of freedom

Multiple R-squared: 0.7662, Adjusted R-squared: 0.7651

F-statistic: 722 on 6 and 1322 DF, p-value: < 2.2e-16

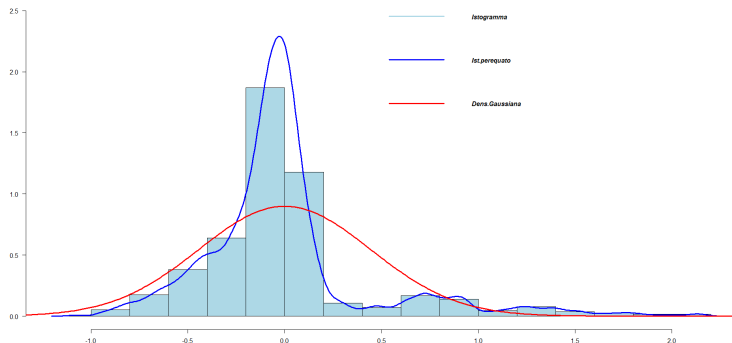
Rappresentazione grafica



Verifica normalità dei residui

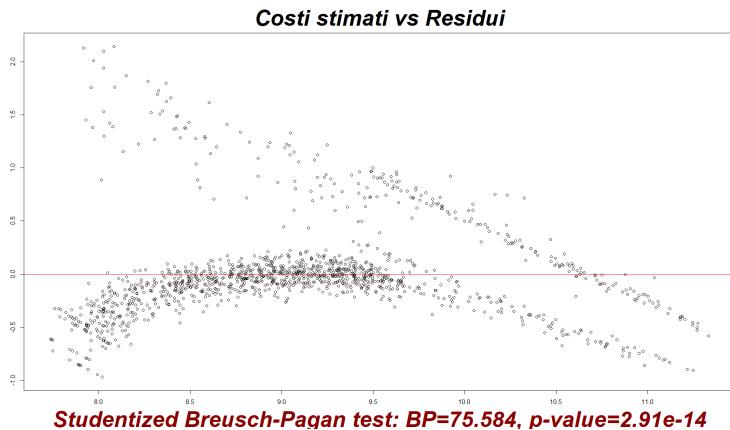
- Conseguentemente alla stima del modello, sono stati effettuati dei test per verificare alcune delle assunzioni sui residui stimati:

Distribuzione dei residui stimati



Shapiro-Wilk normality test: $W=0.84$, $p\text{-value}=2.2e-16$

Verifica omoschedasticità dei residui



Conclusioni

- Avendo dovuto rifiutare buona parte delle ipotesi classiche, sarebbe imprudente utilizzare il modello per scopi predittivi e/o interpretativi.
- La violazione delle ipotesi potrebbe essere dovuta all'assenza di una variabile determinante nella possibile spiegazione dei costi (come il reddito e/o la condizione occupazionale dell'assicurato) oppure al tipo di campionamento effettuato.