



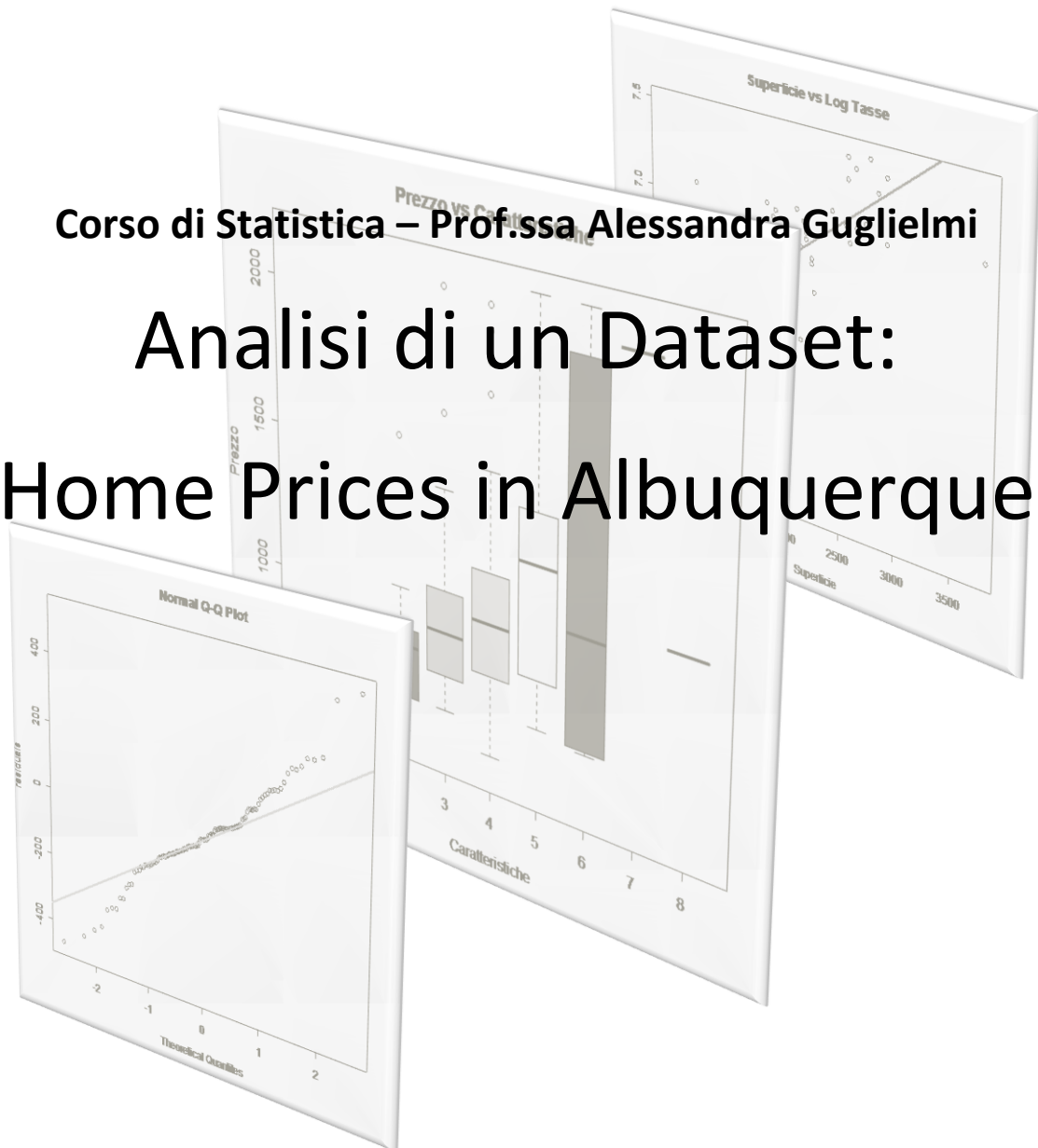
POLITECNICO DI MILANO

IV Facoltà di Ingegneria Industriale

Corso di Laurea in Ingegneria Meccanica

Corso di Statistica – Prof.ssa Alessandra Guglielmi

Analisi di un Dataset: Home Prices in Albuquerque



*Milani Damiano
Minervini Carlo*

Indice

<i>Il dataset.....</i>	<i>3</i>
<i>Problemi e obiettivi dell'analisi</i>	<i>3</i>
<i>Ricostruzione dati mancanti.....</i>	<i>4</i>
<i>Statistica descrittiva.....</i>	<i>4</i>
<i>Regressione lineare semplice.....</i>	<i>5</i>
<i>Regressione lineare multipla e scelta del modello.....</i>	<i>7</i>
<i>Conclusioni.....</i>	<i>8</i>
<i>Appendice.....</i>	<i>9</i>
<i>Indice delle figure.....</i>	<i>9</i>
<i>Output console</i>	<i>25</i>
<i>Script.....</i>	<i>32</i>

Il dataset

I dati presi in esame provengono da un campione raccolto dall' *Albuquerque Board of Realtors* e riguardano gli atti di vendita delle case effettuati nell'omonima città dal 15 febbraio al 30 aprile 1993. Il dataset è composto da un numero di casi pari a 117, per ciascuno dei quali sono state registrate 7 variabili:

- *prezzo*
il prezzo di vendita della casa, in centinaia di dollari
- *superficie*
lo spazio abitabile misurato in piedi quadrati
- *età*
l'età della casa espresso in anni
- *caratteristiche*
contatore del numero di elementi caratteristici presenti nella casa (vengono conteggiati: lavastoviglie, frigorifero, microonde, tritarifiuti, lavatrice, citofono, lucernario, compatta-rifiuti, asciugatrice, accessibilità per disabili, collegamento alla TV via cavo)
- *ne*
indica se la costruzione è posizionata nella zona a Nord-Est (valore 1) della città o meno (valore 0); questa zona di Albuquerque è infatti quella di carattere maggiormente residenziale, con peculiarità più "repubblicane" rispetto al resto della città.
- *angolo*
specifica se l'abitazione è situata in una posizione ad angolo (valore 1) oppure no (valore 0)
- *tasse*
tasse annuali, in dollari

Da evidenziare è il numero non trascurabile di dati mancanti per la variabile *tasse*, ma soprattutto per la variabile *età* (49 valori assenti).

Problema e obiettivi dell'analisi

Il naturale proposito è quello mettere in rilievo le due variabili di maggiore interesse: il prezzo e le tasse della casa; in particolare si vuole spiegare come queste variano in dipendenza alle altre caratteristiche dell'abitazione. Dopo una breve sintesi del campione effettuata per mezzo della statistica descrittiva, si passerà pertanto alla ricerca di modelli di regressione lineare semplice che interpretino queste relazioni. Ulteriore obiettivo sarà quello dell'esame di un modello più completo per il problema, attraverso un modello di regressione multipla.

L'analisi sarà effettuata con l'ausilio del software *R* (www.r-project.org), ma verrà comunque condotta in modo critico, tenendo sempre presente sia la realtà del problema sia i fondamentali concetti statistici.

Ricostruzione dati mancanti

Il dataset è memorizzato su un supporto multimediale codificato in ASCII; è pertanto è facilmente importabile in *R*, se non per il fatto che sono presenti dei dati mancanti evidenziati con il carattere jolly “*”. Ciò causa il mancato riconoscimento delle variabili *età* e *tasse* come di tipo numerico, ed evidenzia la carenza di una modesta percentuale di informazioni. Sebbene sia comunque possibile ricercare un modello di regressione anche con dati mancanti, si è preferito ricostruire su base campionaria la totalità del dataset.

Per questo scopo ci si è serviti della funzione *sample*: per mezzo di questo comando, è stato possibile ripopolare le variabili mancanti, tramite un campionamento casuale (con “reinserimento”) tra i dati già presenti. Più in particolare si sono scansionati i vettori delle variabili in esame con un duplice scopo: in primo luogo per l’identificazione della posizione degli *m* dati mancanti; contemporaneamente si è costruito un nuovo vettore il cui utilizzo sarà demandato alla funzione *sample*. Il contenuto di questo vettore sono gli *n* dati effettivamente noti, dai quali il campionamento sceglierà, con probabilità uniforme, gli *m* nuovi valori candidati a sopperire alle mancanze del dataset. Il passo finale è la sostituzione di questi nuovi dati, di modo da avere un campione completo, senza alcuna perdita di informazioni.

Si presume che i dati mancanti siano dovuti a fattori casuali e non derivanti da carenze specifiche o distorsioni nell’osservazione; ciò è una garanzia che il ripopolamento del dataset, eseguito in maniera similmente *random*, non inficerà la bontà dei modelli ricercati.

Statistica descrittiva

Delle 7 variabili, *ne* e *angolo* sono di tipo categoriale, mentre le restanti sono variabili numeriche intere.

Per un’immediata visualizzazione dei dati si sono realizzati dei grafici per ciascuna delle sette variabili: gli istogrammi e i *boxplot*. Mentre i primi danno un’idea di massima sulla distribuzione dei dati, i *boxplot* permettono di rappresentare i dati evidenziando i cambiamenti tra diversi gruppi di dati. In particolare gli istogrammi graficano la densità per ciascuna classe della variabile considerata. I *boxplot* invece vengono realizzati utilizzando la mediana, il primo e il terzo quantile del campione esaminato; da quest’ultimi due valori viene tracciata una “box” dalla quale partono due code, ciascuna delle quali si estende fino a 1,5 IQR. I valori all’esterno di queste code sono detti *outlier*.

Dagli istogrammi allegati si può notare come nessuna delle variabili evidenzia un particolare andamento noto, quale ad esempio quello di una distribuzione gaussiana; da notare come la variabile *età* sia particolarmente *skewed* verso valori bassi.

Esaminando i *boxplot* invece si denota un marcato cambiamento del *prezzo* in dipendenza alle *caratteristiche*, così pure come un incremento delle *tasse* in corrispondenza della posizione (*ne*).

Regressione lineare semplice

Una regressione lineare semplice esplica una relazione di tipo lineare tra una variabile aleatoria Y , detta risposta, e una variabile numerica x detta regressore; si assume valido il seguente modello

$$Y = \beta_0 + \beta_1 x + \epsilon$$

che poi dovrà essere verificato. In particolare il termine ϵ , indicante l'errore casuale, dovrà distribuirsi secondo una normale con media nulla; questa particolarità verrà studiata con l'analisi dei residui, tecnica che permette di valutare la bontà del modello scelto. Da notare che tale relazione viene detta lineare in quanto lineare nei due parametri β_0 e β_1 .

Lo strumento principale che consente di esaminare un modello di regressione e i parametri da esso derivati è la funzione *lm*. Essa riceve in input le due variabili, indipendente e dipendente, e restituisce il valore dei coefficienti di regressione stimati (stima ai minimi quadrati), la loro significatività e altre statistiche descritte in seguito.

Si vuole quindi indagare una possibile relazione tra le variabili considerate. Per fare ciò si è adottato un approccio metodico per ciascuna delle ipotetiche relazioni. Nella fattispecie saranno valutati:

- la significatività coefficienti: il software R dà indicazioni sulla validità dei coefficienti di regressione stimati, a seconda del p-value di un test che ha come ipotesi H_0 che i coefficienti siano nulli.
- R-quadro: è il coefficiente di determinazione, parametro che varia tra 0 e 1, di fondamentale importanza per valutare la percentuale della variabilità totale spiegata dal modello lineare.
- *F statistic*: statistica test che ha come ipotesi nulla il caso in cui tutti i coefficienti di regressione abbiano contemporaneamente valore 0.
- analisi dei residui: si vuole verificare la normalità dei residui derivanti dal modello ipotizzato, come conseguenza della definizione di modello lineare; per fare ciò si utilizzano 2 tecniche: una grafica, il *qq-plot*, che rappresenta i quantili della distribuzione; se questi si dispongono lungo una retta allora la distribuzione dei residui è normale; la seconda tecnica è di tipo numerico ed è basata sullo *Shapiro-Wilk test*, un test di adattamento che ha come ipotesi nulla la normalità dei residui. Ulteriore caratteristica dei residui è che essi devono avere la medesima varianza; ci si serve a tal scopo di un riscontro grafico, lo *scatterplot* dei residui, che dovrebbe assumere una dispersione a "nuvola di punti" senza particolari tendenze di forma (ciò avalla l'ipotesi di omoschedasticità).

Nel caso in cui sia dubbia la bontà del modello, a causa della non normalità dei residui, si procederà alla trasformazione dei dati, dapprima per mezzo di una trasformazione logaritmica, e in secondo luogo tramite una *boxcox*. In particolare R sceglie la migliore trasformazione di questa famiglia rappresentata dalla generica trasformazione

$$T(X_i, \lambda) = \frac{X_i^\lambda - 1}{\lambda}, \quad \lambda \in [-2, +2]$$

ovvero il software seleziona il valore di λ tale che il p-value dello *Shapiro test* sia il più piccolo. Si noti che per $\lambda \rightarrow 0$ ci si riconduce al caso di trasformazione logaritmica.

Per la scelta delle coppie di variabili tra cui potrebbe esistere un rapporto di regressione lineare, ci si è serviti del coefficiente di correlazione lineare tra le stesse, per avere, in prima approssimazione, un'idea sui legami esistenti. Unitamente a ciò e alla correlazione suggerita dal buonsenso, si sono analizzate le seguenti coppie:

1. Regressione prezzo-superficie

- significatività coefficienti: l'intercetta è poco significativa, come facilmente prevedibile in quanto a superficie nulla corrisponderà un prezzo nullo; il coefficiente del regressore risulta invece molto significativo.
- R-quadro: pari a 0,71 è indice che il modello spiega bene i dati.
- *F statistic*: ha p-value molto basso, non c'è evidenza sperimentale che i coefficienti siano entrambi pari a 0.
- analisi dei residui: lo *Shapiro-test* è bassissimo, pertanto i residui non seguono un andamento normale; ciò è anche confermato dall'andamento del *qq-plot* dove si evince che la distribuzione ha code più pesanti della normale.
- trasformazioni: nemmeno trasformando i dati con una *boxcox* o logaritmica si sono ottenuti residui distribuiti normalmente.

2. Regressione prezzo-caratteristiche

- significatività coefficienti: entrambi i coefficienti sono significativi.
- R-quadro: non è molto elevato.
- *F statistic*: ha p-value molto basso, non c'è evidenza sperimentale che i coefficienti siano entrambi pari a 0.
- analisi dei residui: il p-value dello *Shapiro-test* è basso; il *qq-plot* invece denota una distribuzione non gaussiana.
- trasformazioni: c'è un sensibile miglioramento del modello in caso di trasformazione *boxcox* con $\lambda = -0.8$; in questo modo i residui hanno un andamento normale (p-value dello *Shapiro-test* pari a 0.32).

3. Regressione tasse-superficie

- significatività coefficienti: per lo stesso motivo esposto in precedenza, l'intercetta è poco significativa; il coefficiente del regressore è invece buono.
- R-quadro: è pari a 0,63.
- *F statistic*: ha p-value molto basso, non c'è evidenza sperimentale che i coefficienti siano entrambi pari a 0.
- analisi dei residui: il p-value dello *Shapiro-test* è basso; il *qq-plot* invece denota una distribuzione con code più pesanti rispetto a una gaussiana.
- trasformazioni: l'ipotesi di modello lineare si fa ancora più critica nel caso di trasformazione dei dati.

4. Regressione tasse-prezzo

- significatività coefficienti: intercetta non significativa a differenza del coefficiente angolare.
- R-quadro: il coefficiente di determinazione è abbastanza elevato, ha un valore di 0,7.
- *F statistic*: ha p-value molto basso, non c'è evidenza sperimentale che i coefficienti siano entrambi pari a 0.

- analisi dei residui: il valore del p-value dello *Shapiro-test* è molto basso, fatto che ci fa rifiutare l'ipotesi che i residui siano distribuiti normalmente; il *qq-plot* invece denota una distribuzione con code più pesanti rispetto a una gaussiana.
- trasformazioni: lieve miglioramento della normalità dei residui, sebbene non si possa comunque validare il modello lineare.

Regressione lineare multipla e scelta del modello

Un'estensione del modello lineare semplice è quella di regressione lineare multipla; in questo caso la variabile aleatoria Y è spiegata da più variabili indipendenti x_i . La relazione rimane comunque di tipo lineare in quanto è lineare nei parametri β_i :

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$$

Analogamente al caso semplice, sempre con l'ausilio del comando *lm*, si studierà la possibile presenza di un modello lineare multiplo tra più variabili, esaminando la significatività dei coefficienti, il coefficiente di determinazione, la *F statistic* e l'analisi dei residui.

Particolare attenzione sarà posta su un'eventuale multicollinearità; la bassa significatività per tutti i parametri e la *F statistic* con p-value basso possono essere sintomo di mutua dipendenza tra i regressori. Dopo un'analisi per via grafica effettuabile tramite il comando *pairs*, si utilizzerà la funzione *vif* (*variance inflation factor*), che valuta il coefficiente di determinazione di una relazione del tipo

$$X_1 = \beta_0 + \beta_1 x_2 + \dots + \beta_k x_k + \epsilon$$

In particolare se il valore

$$vif(x_i) = \frac{1}{1 - R^2_{x_i}}$$

risulta essere maggiore di 10, è presente multicollinearità.

Nel caso in esame, in seguito a una sospetta multicollinearità tra *tasse* e *prezzo* dovuta al particolare andamento dei diagrammi output di *pairs*, una più approfondita analisi con la funzione *vif* non ha evidenziato alcuna collinearità tra le variabili. Pertanto si può procedere ad un'analisi di regressione multipla senza incorrere in errori dovuti a ridondanza di informazioni.

Dopo una prima regressione tra *prezzo* e la totalità delle rimanenti variabili, da un'iniziale analisi dei risultati ottenuti si è notato come non tutte le variabili considerate siano significative per il modello multiplo. Per la scelta di un modello migliore ci si è serviti della funzione *step*, che è basata sull'AIC (*Akaike's information criterion*), la quale fa una cernita delle variabili che contribuiscono a dare maggiore significatività al modello. Da questa analisi è pertanto derivato un modello che contempla come regressori *superficie*, *angolo* e *tasse*.

È stata inoltre effettuata un'analisi sulle variabili di tipo binario (*dummy variable*) *angolo* e *ne*, e la loro influenza sulla relazione *prezzo-tasse*, ipotizzando una dipendenza di questa dalla posizione della casa. Tuttavia non è stata riscontrata una demarcata separazione tra le regressioni per i due valori delle *dummy variables* (come si vede immediatamente dagli *scatter plot* allegati).

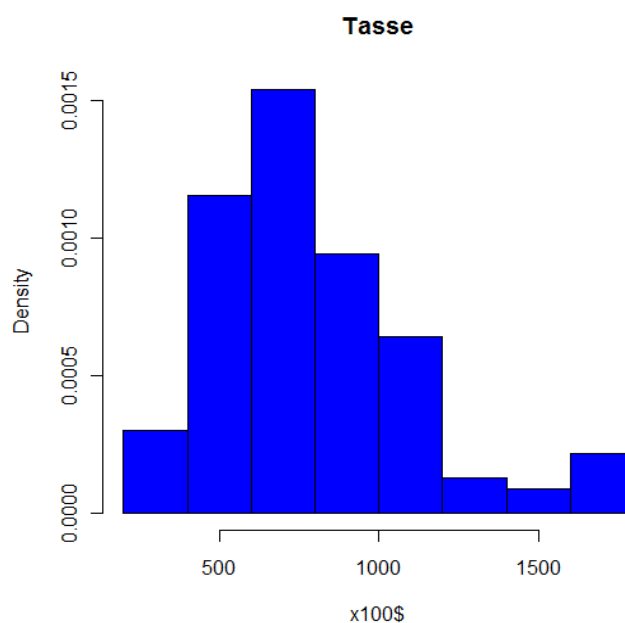
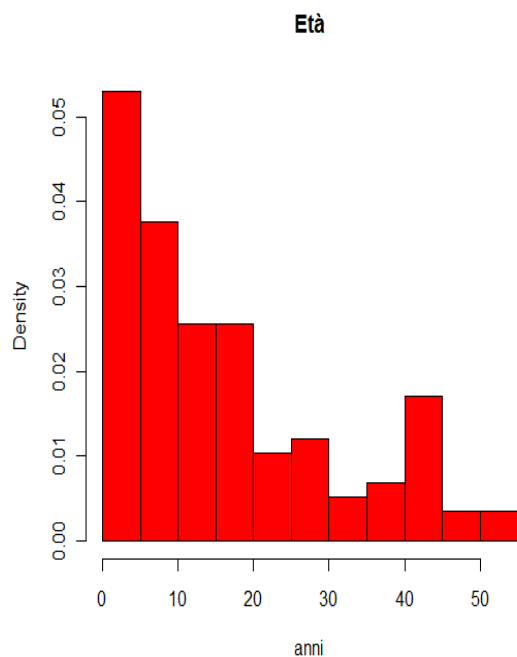
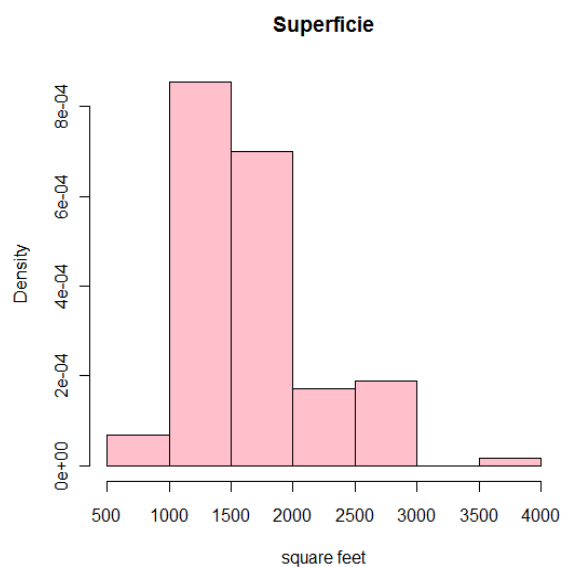
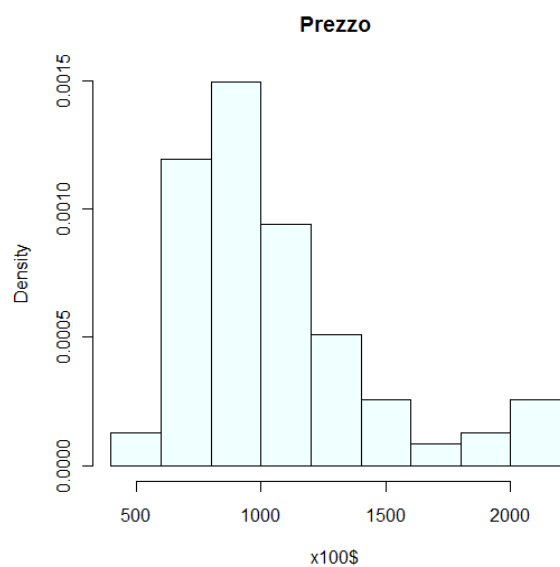
Conclusioni

In definitiva si può desumere che il modello lineare non è un buon modello che descrive i legami tra le variabili del dataset. Questa è la conclusione a cui si perviene a causa della non normalità dei residui, anche nel caso di trasformazione dei dati. La regressione lineare semplice è solamente riscontrabile nel caso *prezzo-caratteristiche*. Non si può nemmeno validare il modello di regressione multipla, dato che ancora una volta l'analisi dei residui rigetta l'ipotesi di linearità.

Appendice

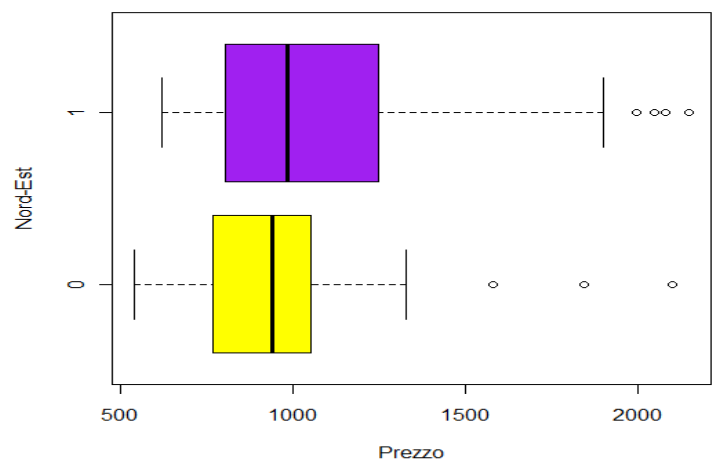
Indice delle figure

STATISTICA DESCRITTIVA - Istogrammi

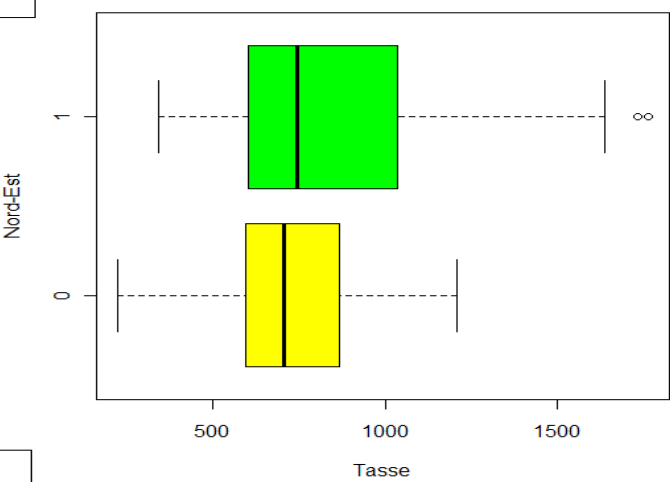


Boxplot

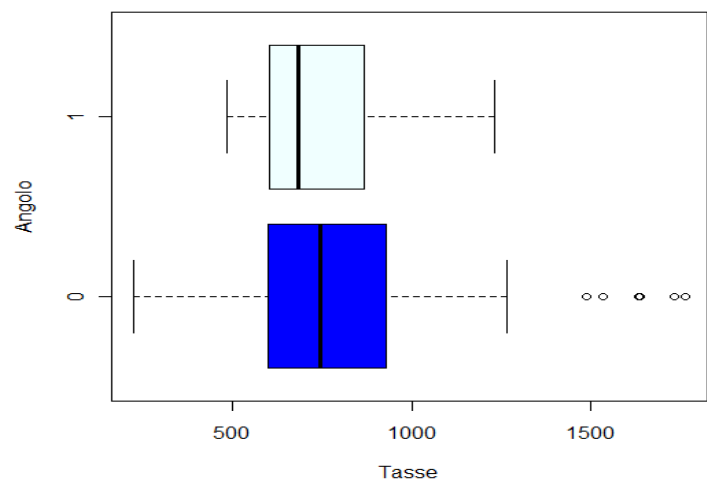
Prezzo vs Nord-Est



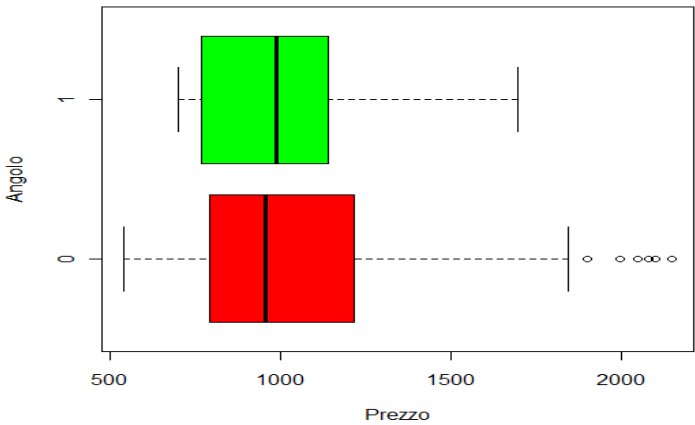
Tasse vs Nord-Est



Tasse vs Angolo

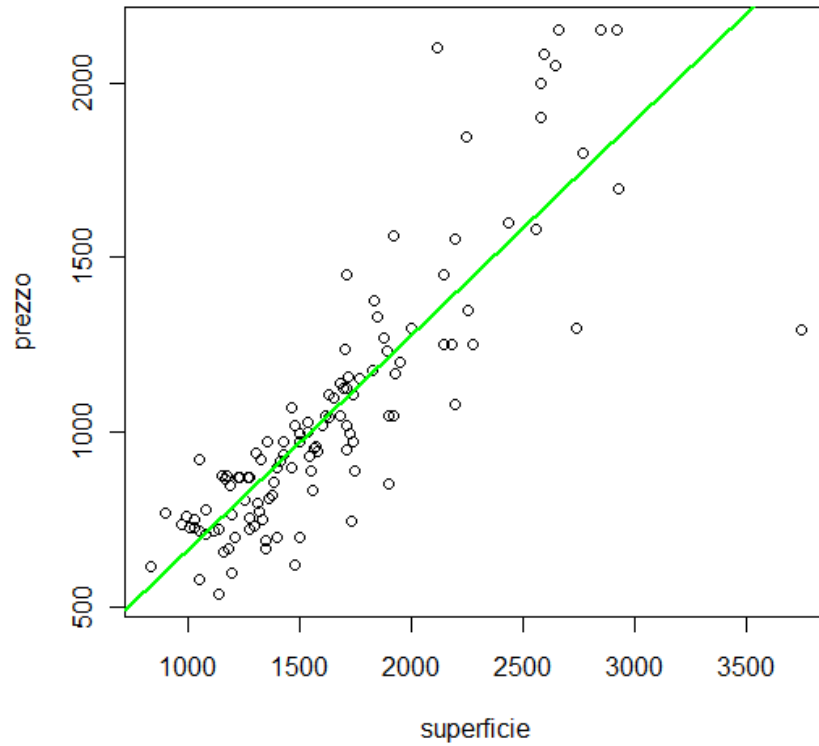


Prezzo vs Angolo

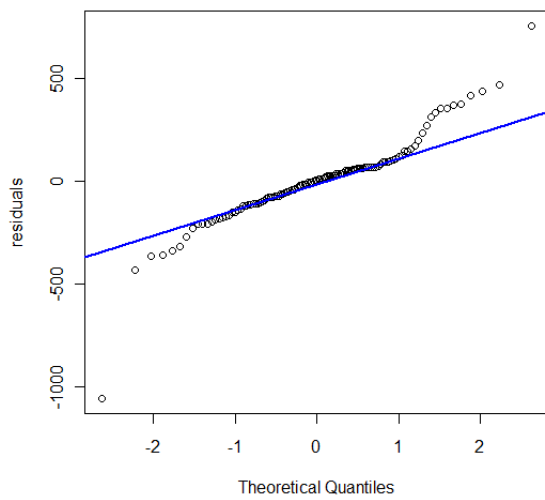


REGRESSIONE LINEARE SEMPLICE

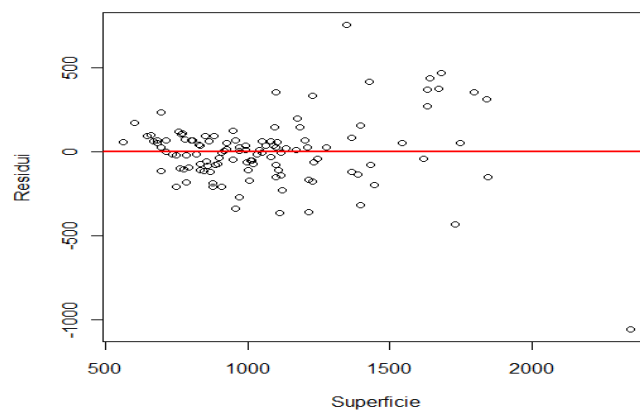
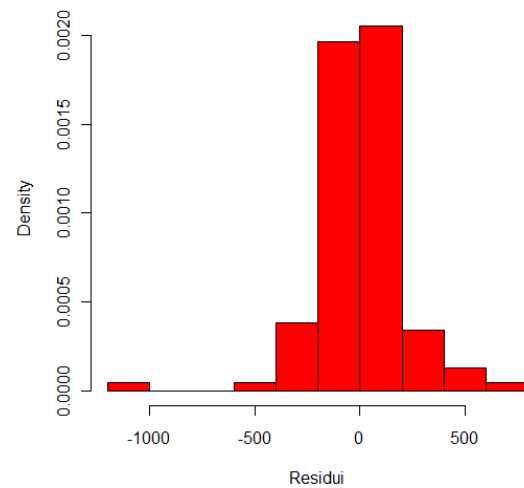
Regressione prezzo-superficie



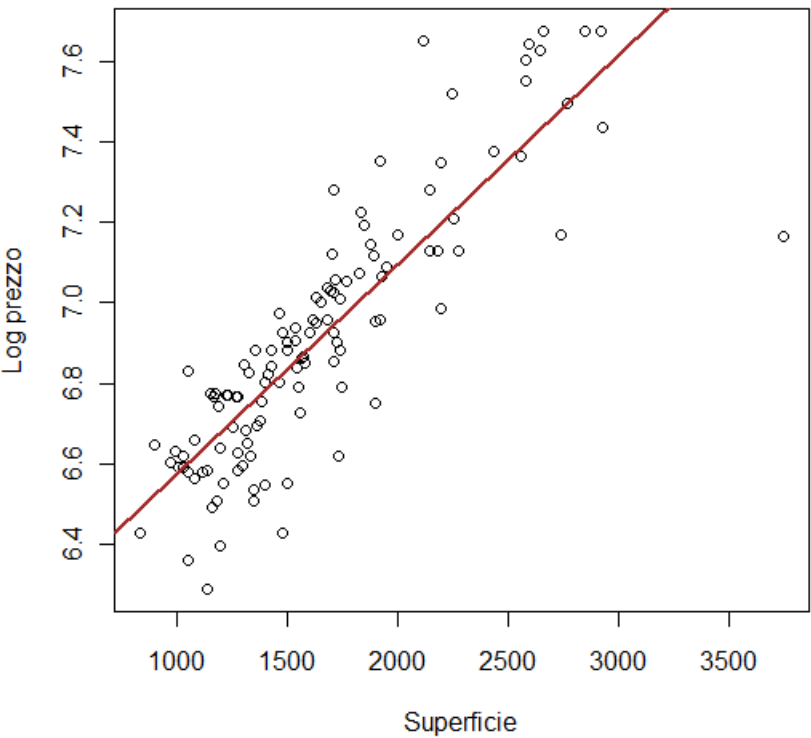
Normal Q-Q Plot



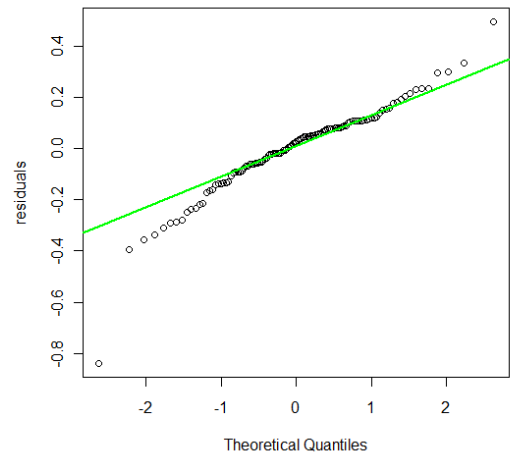
istogramma residui regr. prezzo-superficie



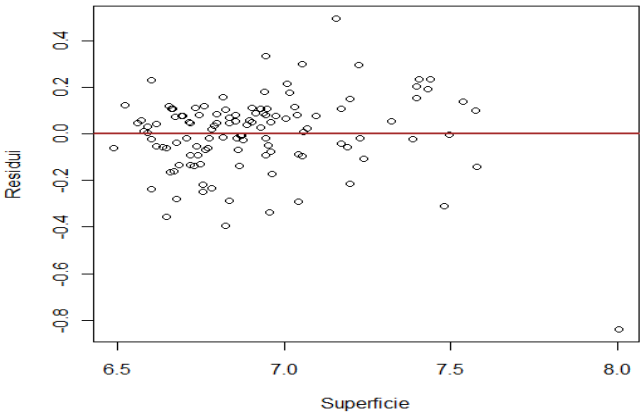
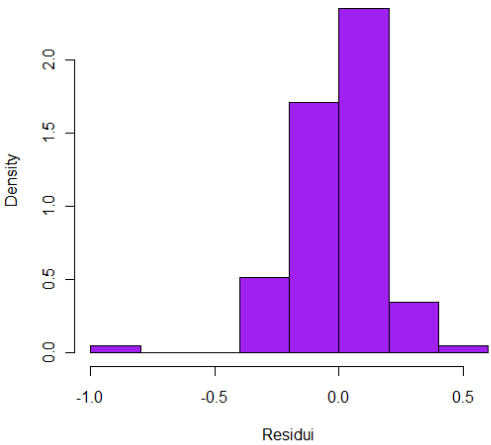
Regressione prezzo-superficie – Trasformazione logaritmica



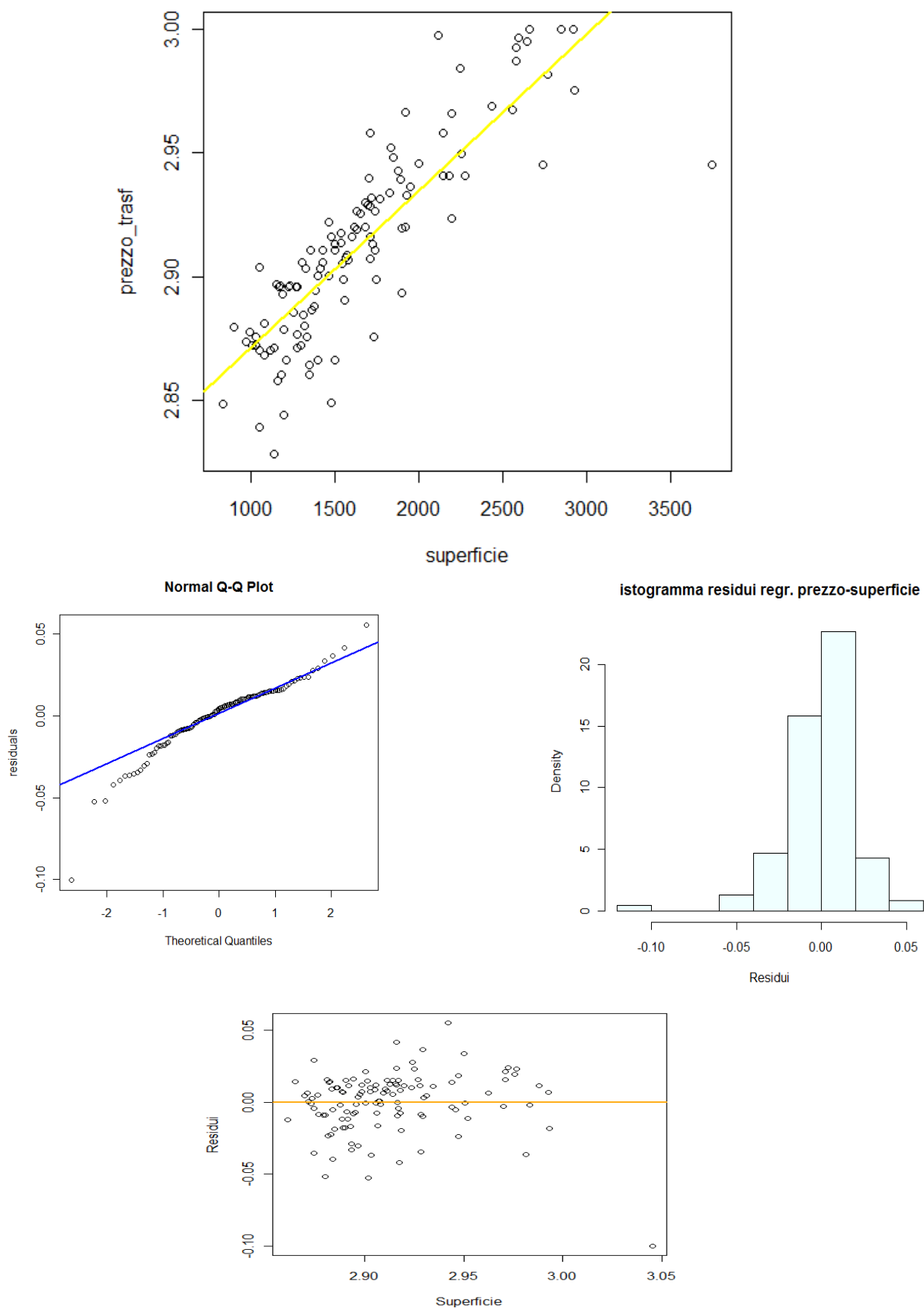
Normal Q-Q Plot



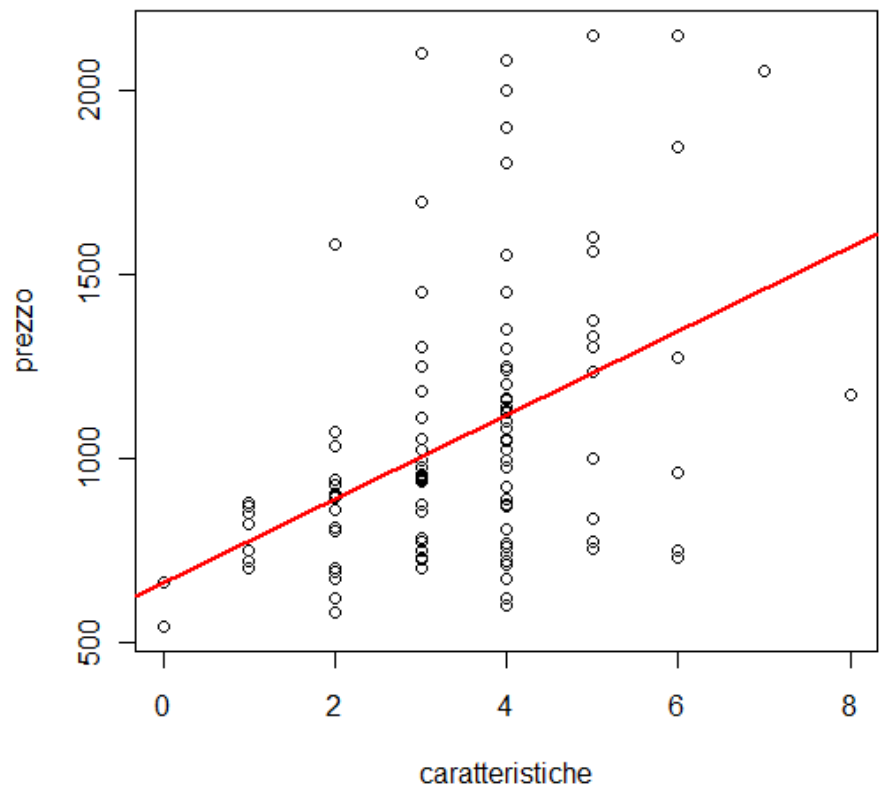
istogramma residui regr. log(prezzo)-superficie



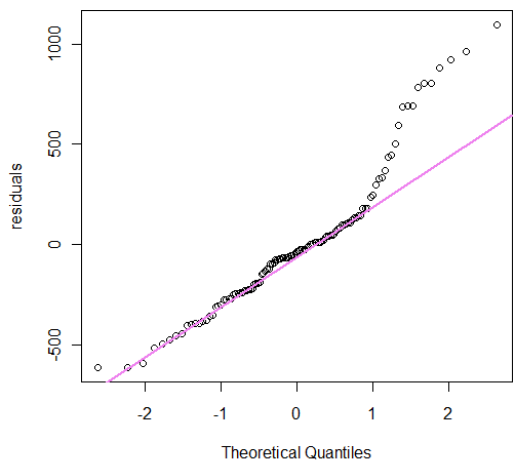
Regressione prezzo-superficie – Trasformazione boxcox



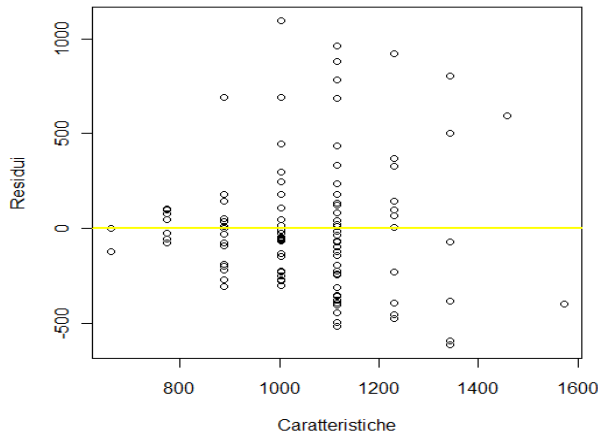
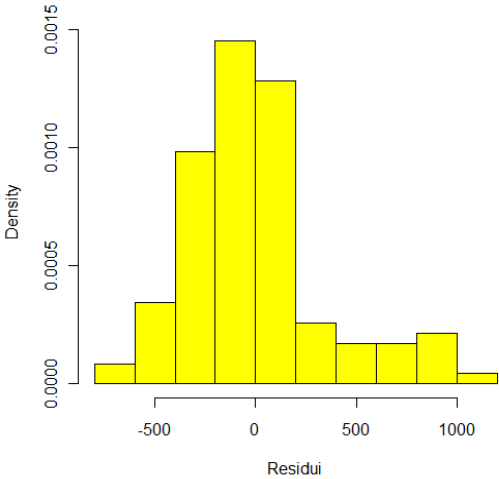
Regressione prezzo-caratteristiche



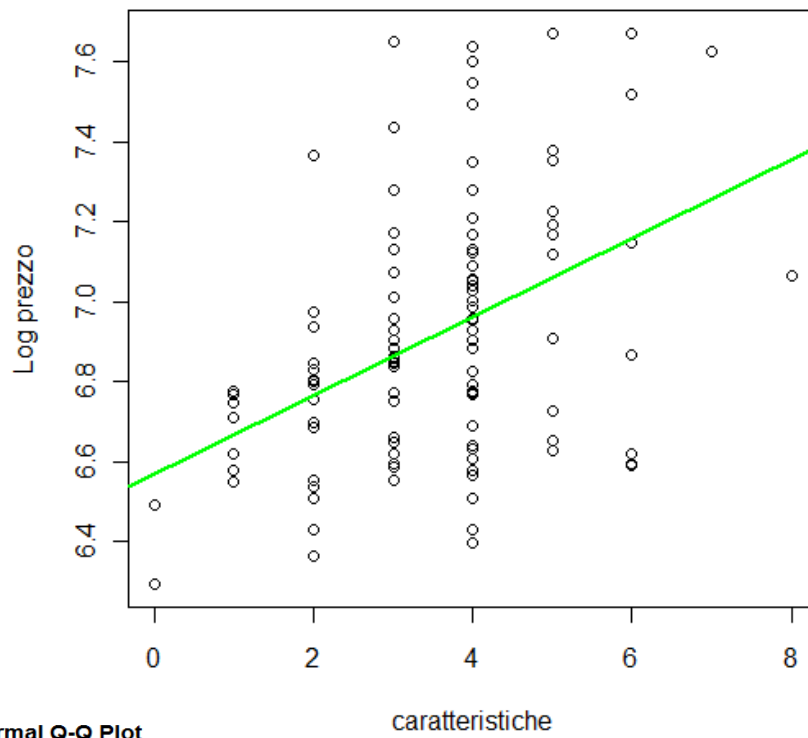
Normal Q-Q Plot



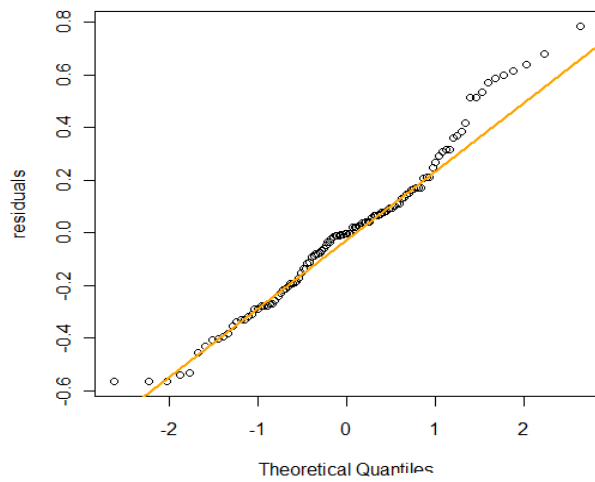
istogramma residui regr. prezzo-caratteristiche



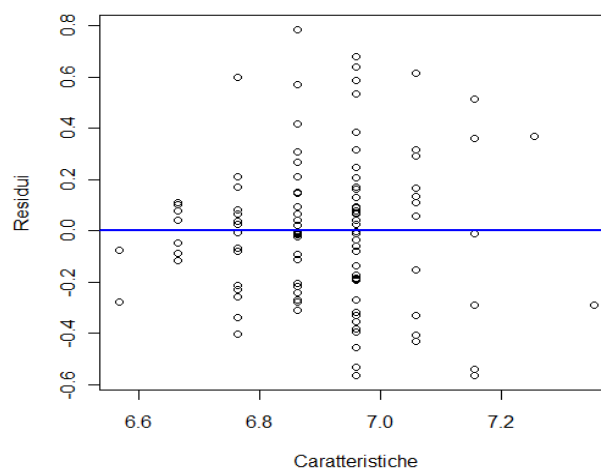
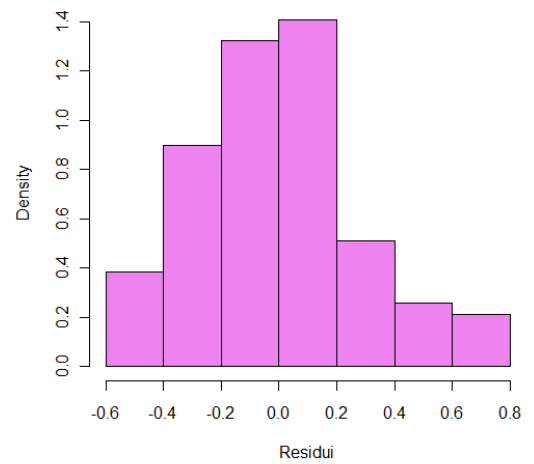
Regressione prezzo-caratteristiche – Trasformazione logaritmica



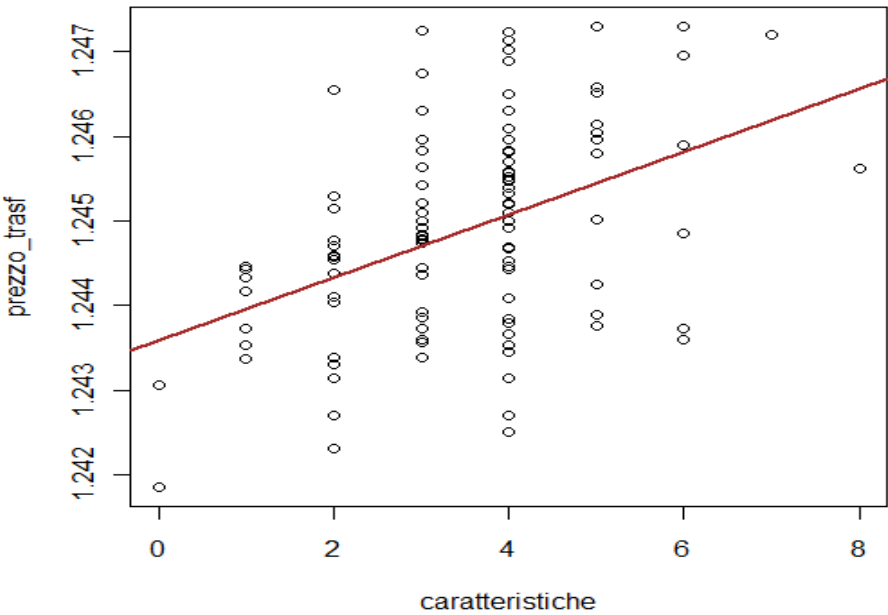
Normal Q-Q Plot



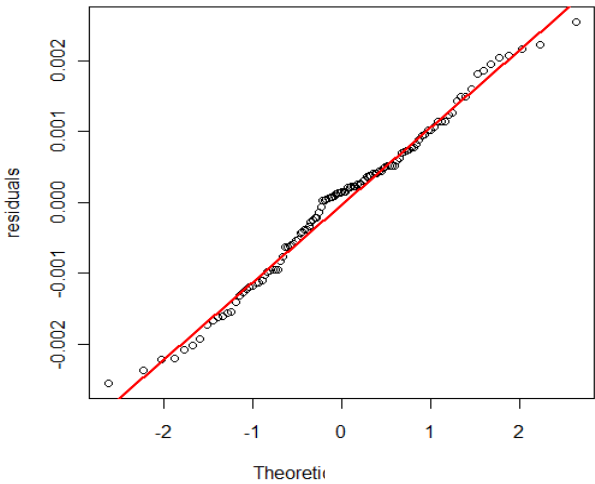
istogramma residui regr. log(prezzo)-caratteristiche



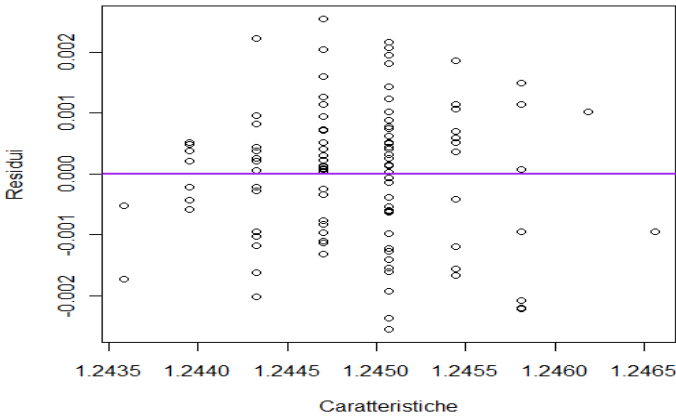
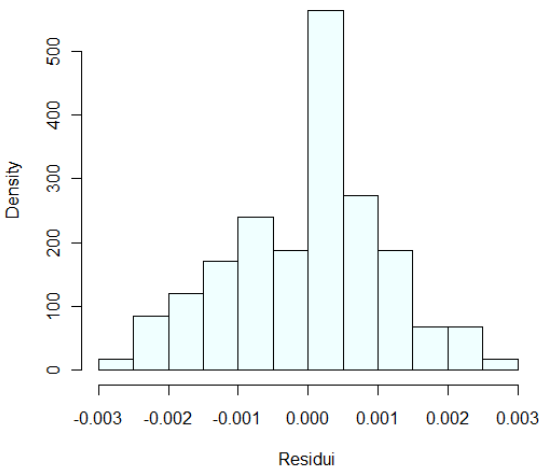
Regressione prezzo-caratteristiche – Trasformazione boxcox



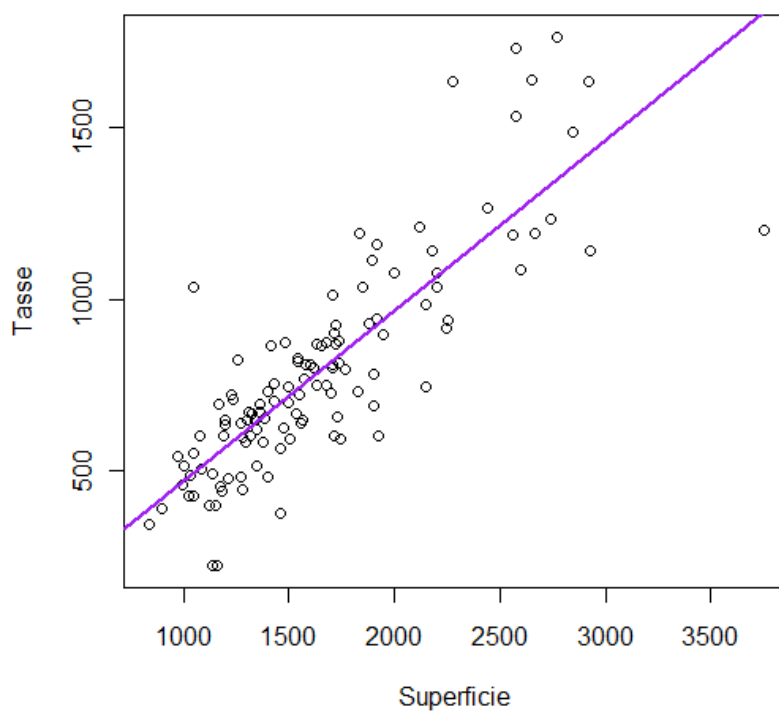
Normal Q-Q Plot



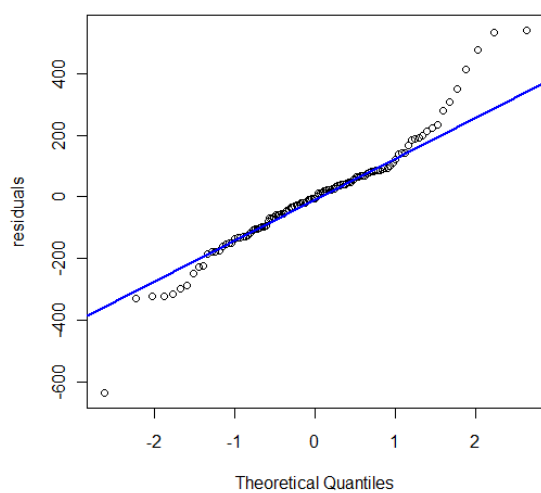
istogramma residui regr. prezzo-caratteristiche



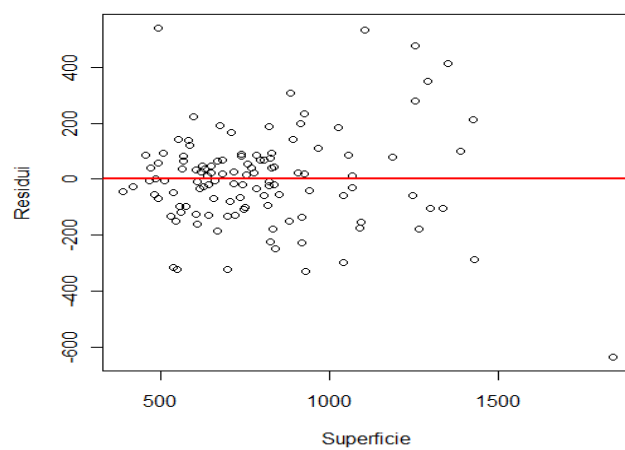
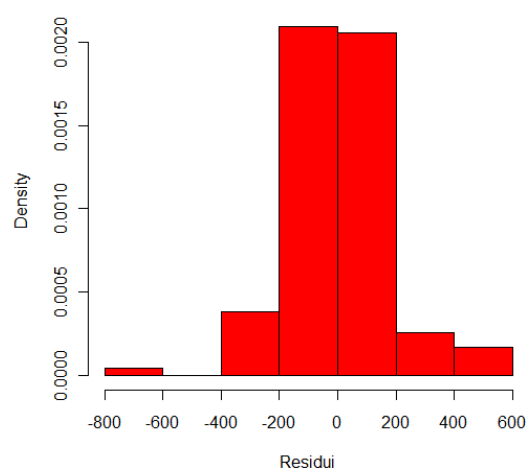
Regressione tasse-superficie



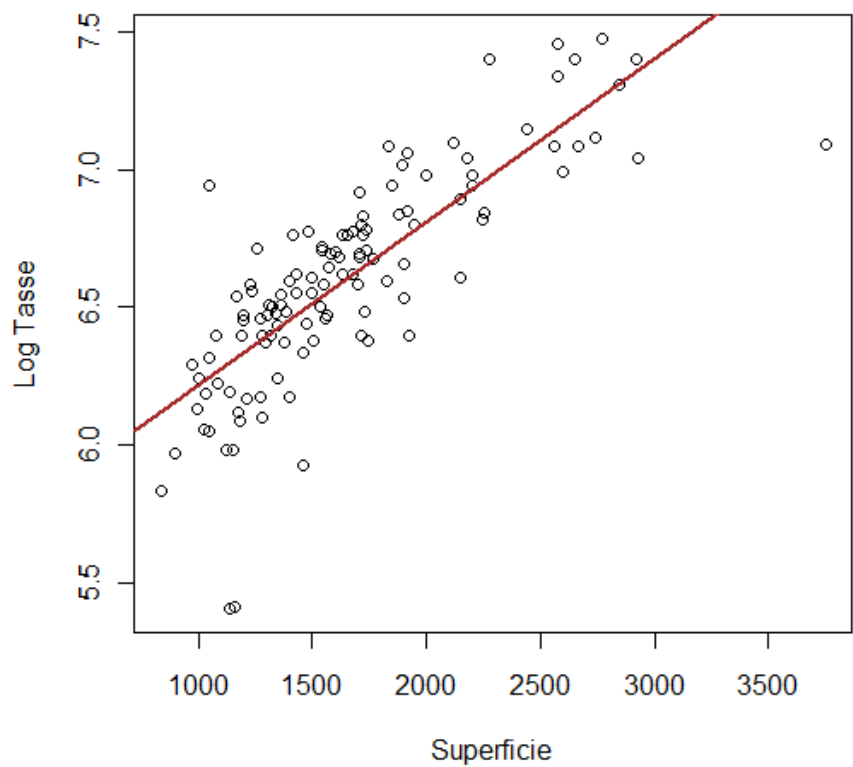
Normal Q-Q Plot



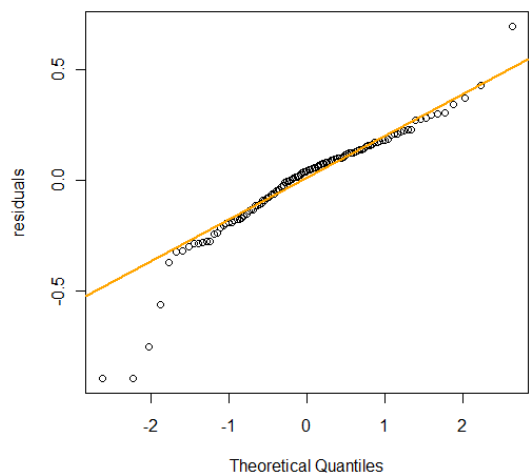
istogramma residui regr. tasse-superficie



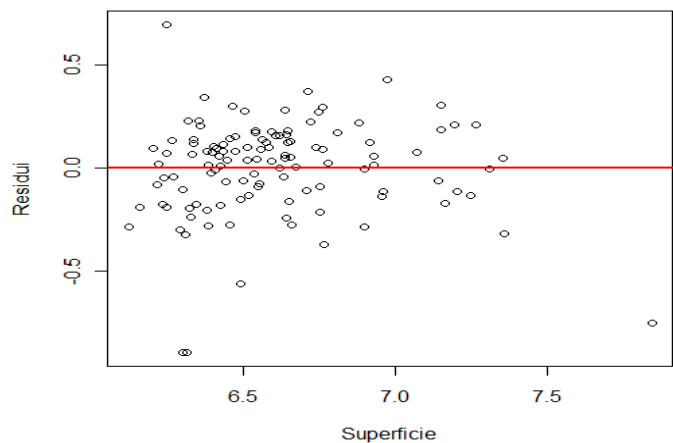
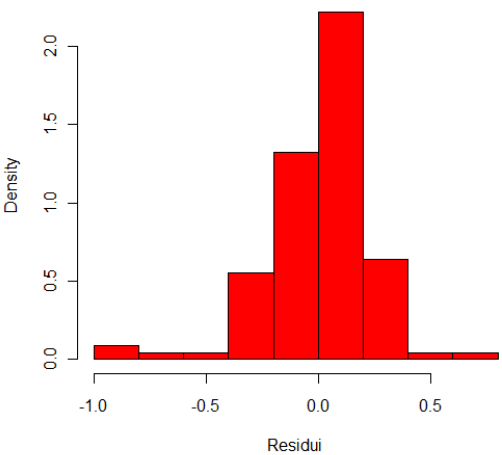
Regressione tasse-superficie – Trasformazione Logaritmica



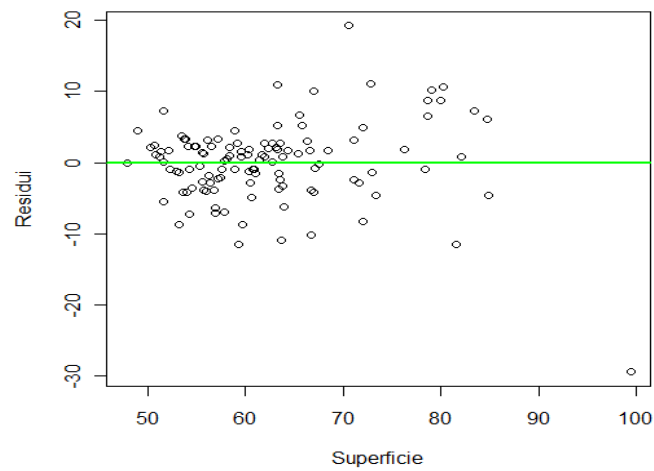
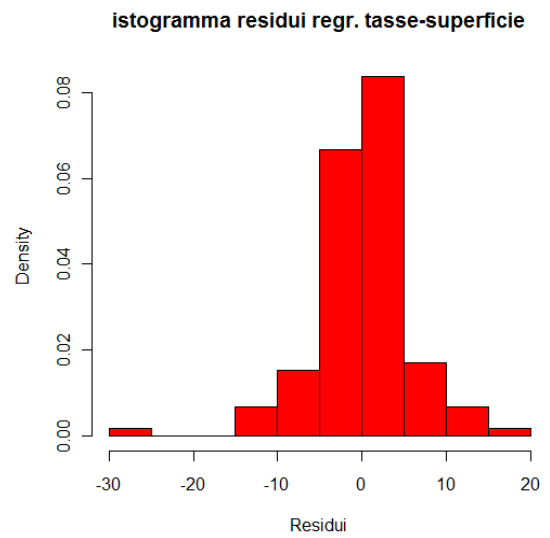
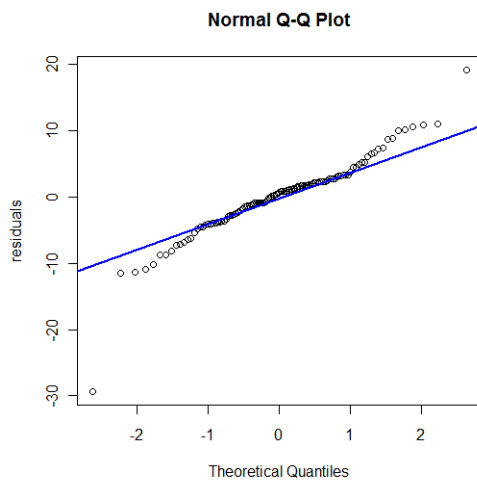
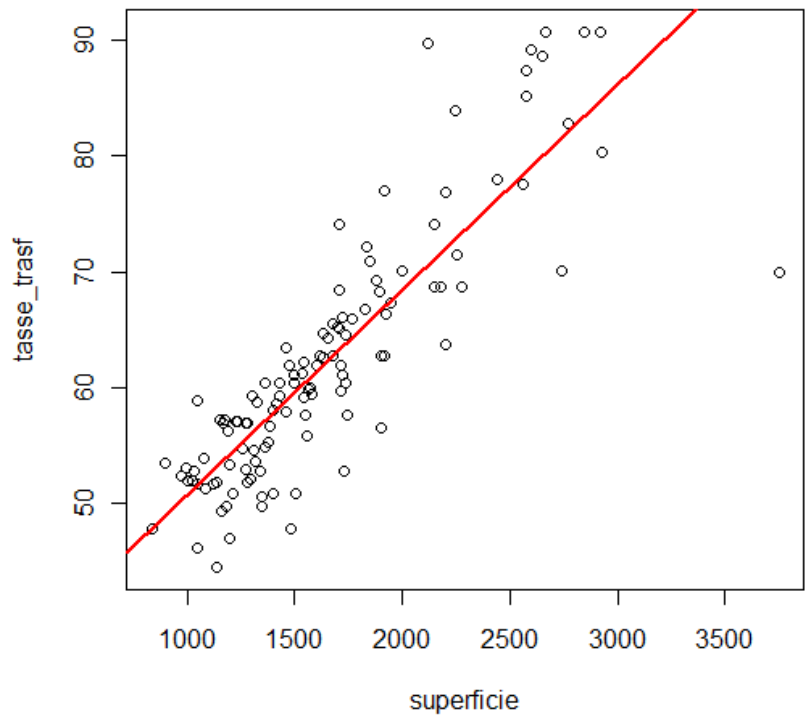
Normal Q-Q Plot



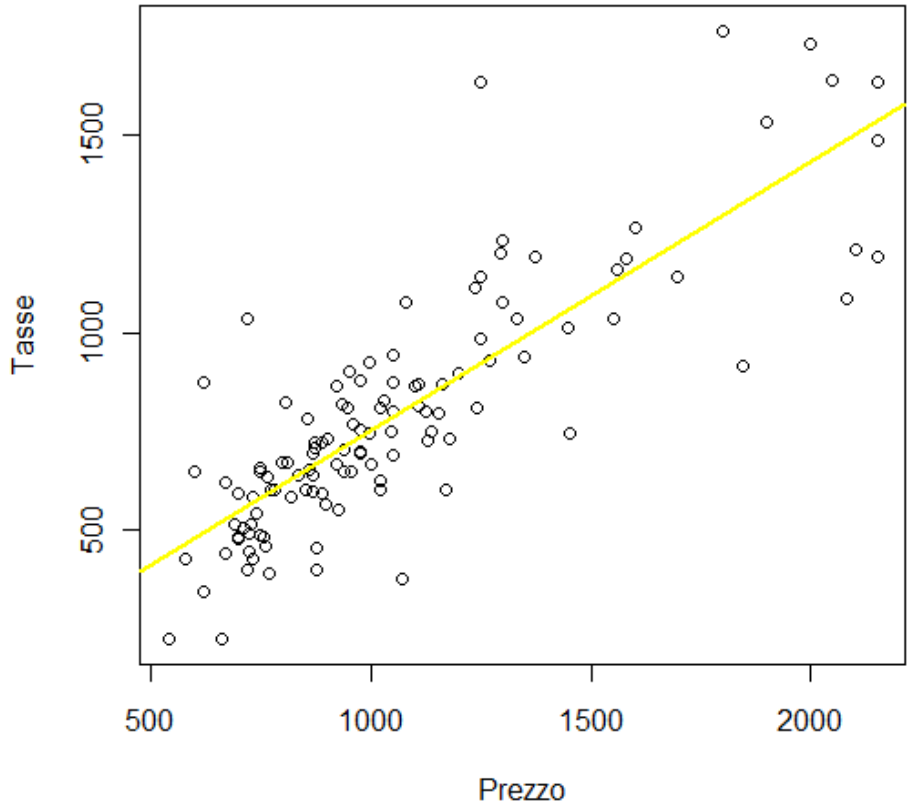
istogramma residui regr. log(tasse)-superficie



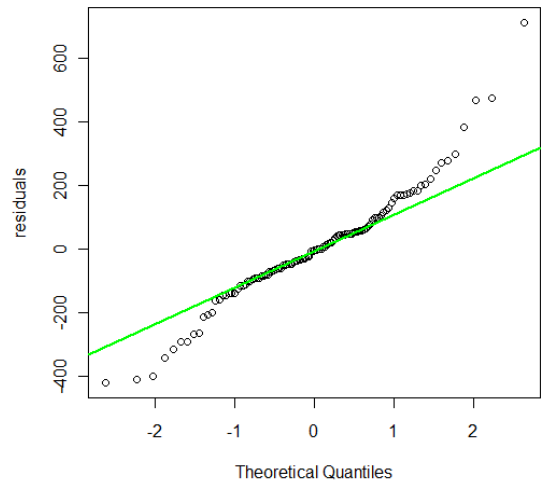
Regressione tasse-superficie – Trasformazione BoxCox



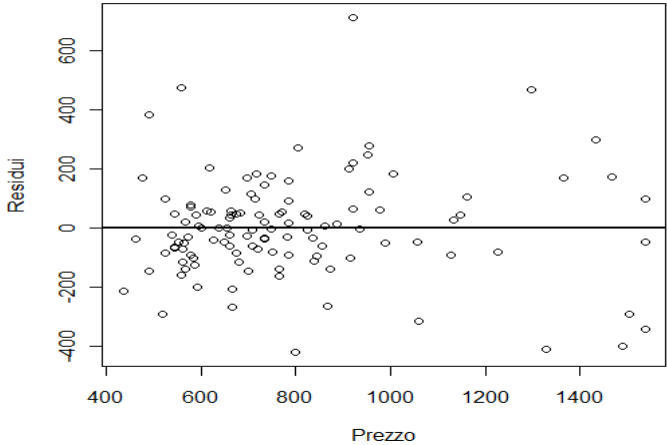
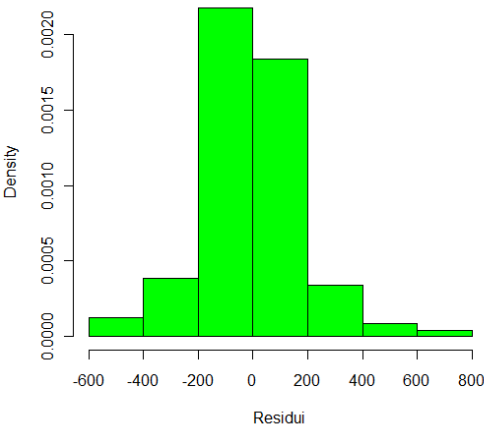
Regressione tasse-prezzo



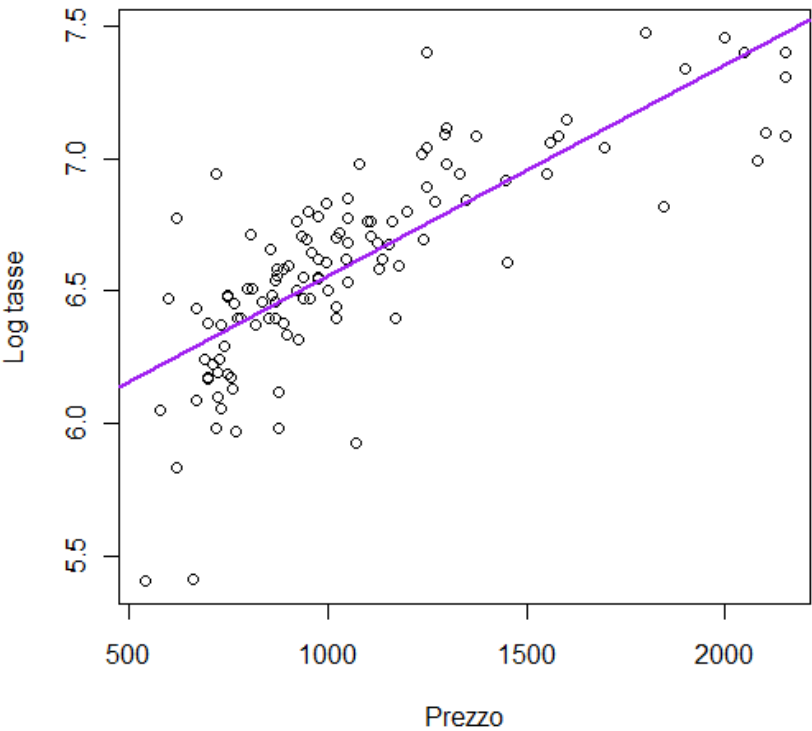
Normal Q-Q Plot



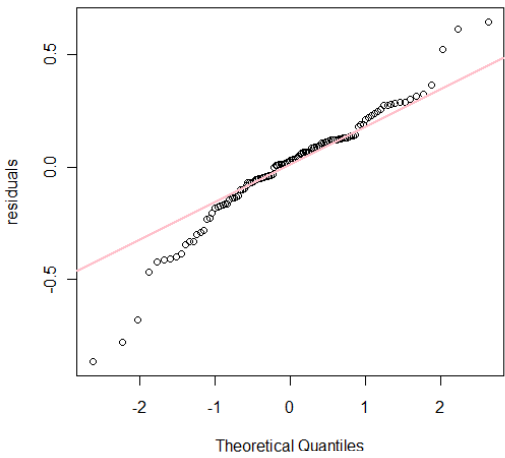
istogramma residui regr. tasse-prezzo



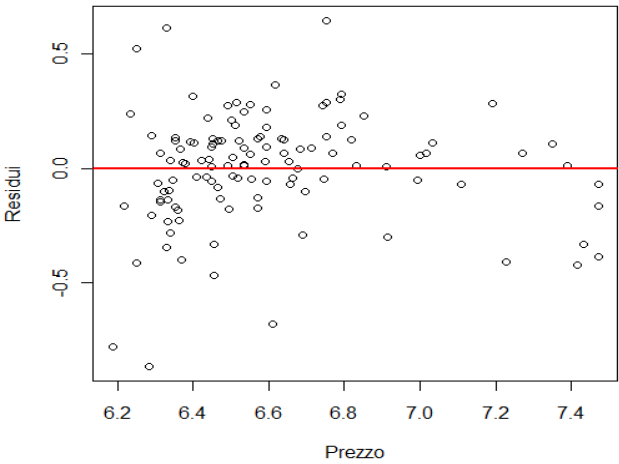
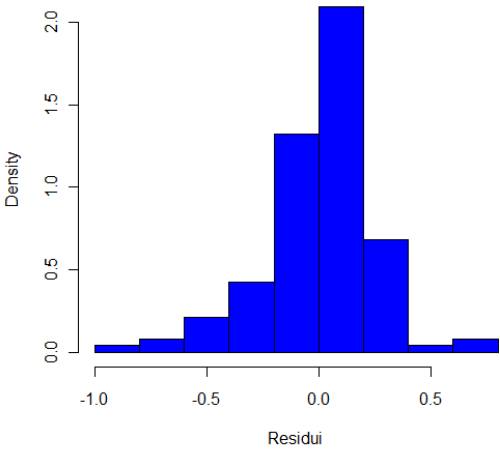
Regressione tasse-prezzo – Trasformazione logaritmica



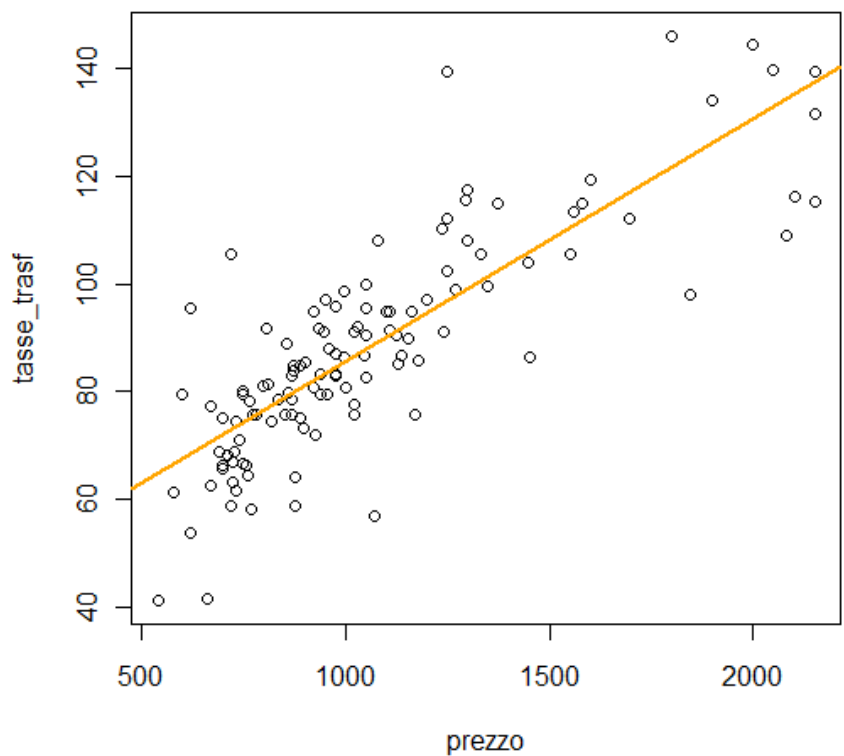
Normal Q-Q Plot



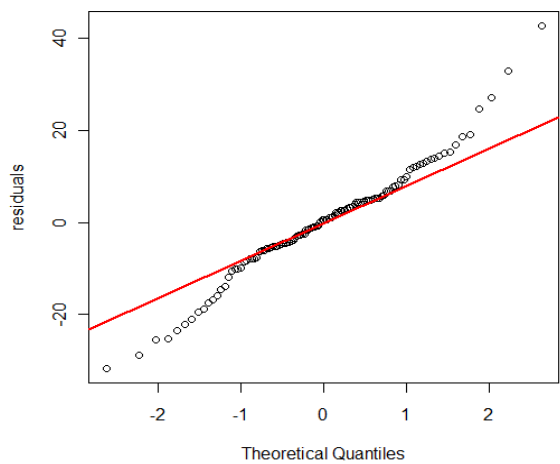
istogramma residui regr. log(tasse)-prezzo



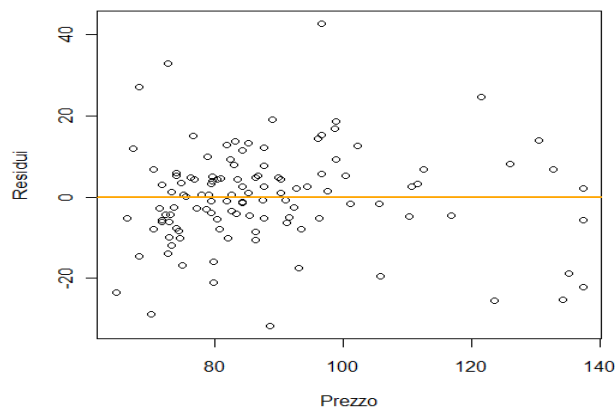
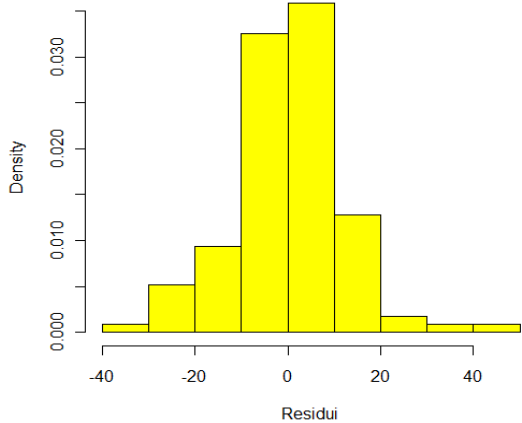
Regressione tasse-prezzo – Trasformazione boxcox



Normal Q-Q Plot

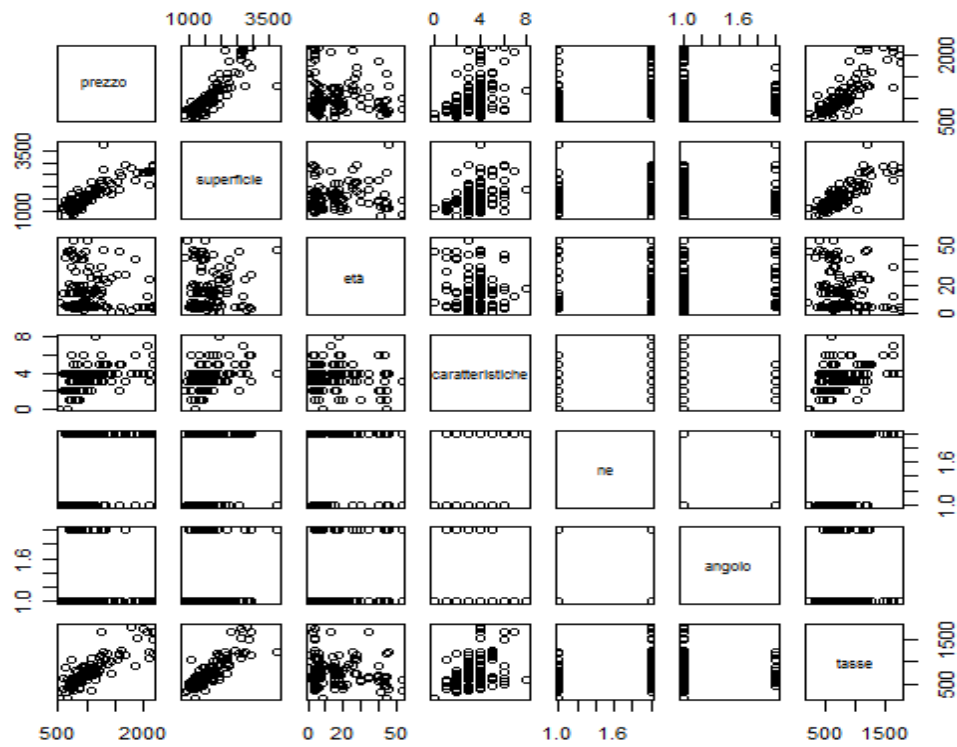


istogramma residui regr. tasse-prezzo

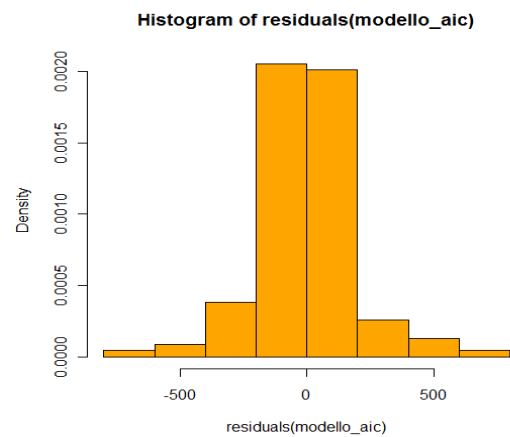
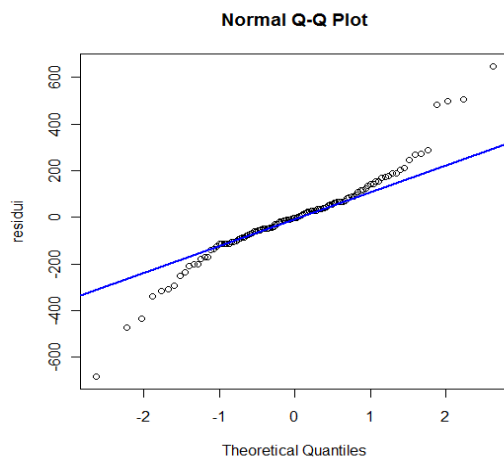


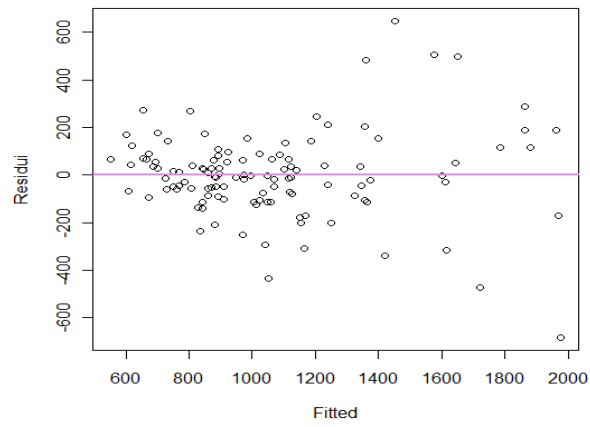
REGRESSIONE MULTIPLA

Multicollinearità

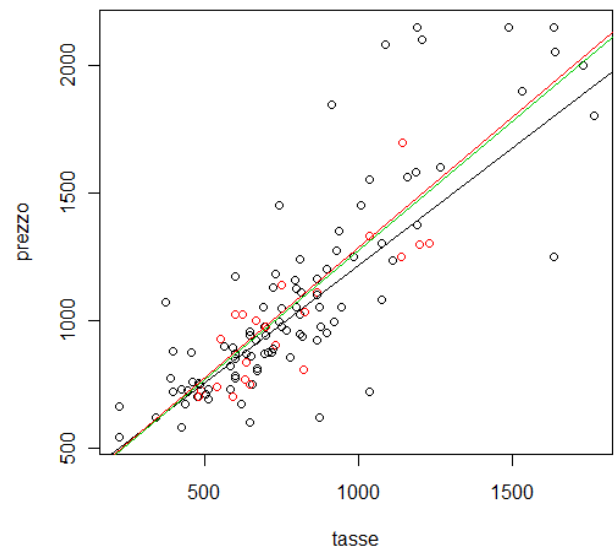
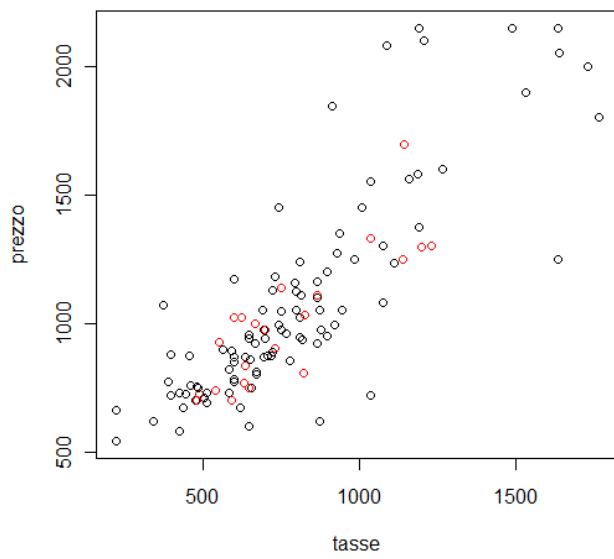


Modello multiplo





DUMMY VARIABLES



Output Console

Coefficiente di correlazione:

	prezzo	superficie	età	caratteristiche	ne
prezzo	1.0000000	0.84479510	-0.12666591	0.42027250	0.16784024
superficie	0.8447951	1.00000000	0.00345716	0.39492498	0.14502997
età	-0.1266659	0.00345716	1.00000000	-0.11128524	0.12458196
caratteristiche	0.4202725	0.39492498	-0.11128524	1.00000000	0.19001561
ne	0.1678402	0.14502997	0.12458196	0.19001561	1.00000000
angolo	-0.0792926	0.04052797	0.10643832	-0.04154692	-0.07733603
tasse	0.8327165	0.83466853	-0.19197714	0.41123117	0.16885474

	angolo	tasse
prezzo	-0.07929260	0.8327165
superficie	0.04052797	0.8346685
età	0.10643832	-0.1919771
caratteristiche	-0.04154692	0.4112312
ne	-0.07733603	0.1688547
angolo	1.00000000	-0.0384415
tasse	-0.03844150	1.0000000

```
#=====#  
# REGRESSIONE PREZZO-SUPERFICIE #  
#=====#
```

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercetta)	47.81931	62.85482	0.761	0.448	
superficie	0.61367	0.03625	16.931	<2e-16	***

Multiple R-squared: 0.7137, Adjusted R-squared: 0.7112

F-statistic: 286.6 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept)	-76.6839932	172.322606
superficie	0.5418706	0.685463

Analisi dei residui

Shapiro-Wilk normality test

W = 0.8948, p-value = 1.436e-07

Trasformazione logaritmica dei dati:

Regressione $\log(\text{prezzo}) \sim \text{superficie}$

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.055e	5.179e-02	116.90	<2e-16	***
superficie	5.200e-04	2.987e-05	17.41	<2e-16	***

Multiple R-squared: 0.7249, Adjusted R-squared: 0.7225

F-statistic: 303.1 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept)	5.951983814	6.1571714528
superficie	0.000460785	0.0005791087

Analisi dei residui

Shapiro-Wilk normality test

W = 0.9277, p-value = 8.82e-06

Trasformazione Boxcox dei dati:

lambda= -0.3

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.808	6.452e-03	435.24	<2e-16	***
superficie	6.327e-05	3.720e-06	17.01	<2e-16	***

Multiple R-squared: 0.7155, Adjusted R-squared: 0.713

F-statistic: 289.2 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept)	2.795239	2.820798e+00
superficie	5.590396e-05	7.064293e-05

Analisi dei residui

Shapiro-Wilk normality test

W = 0.925, p-value = 6.08e-06

```
#===== #
#REGRESSIONE PREZZO-CARATTERISTICHE #
#===== #
```

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	661.17	86.97	7.602	8.58e-12	***
caratteristiche	113.76	22.90	4.967	2.39e-06	***

Multiple R-squared: 0.1766, Adjusted R-squared: 0.1695
F-statistic: 24.67 on 1 and 115 DF, p-value: 2.388e-06

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept)	488.90322	833.4451
caratteristiche	68.39161	159.1271

Analisi dei residui
Shapiro-Wilk normality test
W = 0.9204, p-value = 3.269e-06

Trasformazione logaritmica dei dati:

Regressione log(prezzo) ~caratteristiche

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.56801	0.07269	90.351	< 2e-16	***
Caratteristiche	0.09816	0.01914	5.127	1.20e-06	***

Multiple R-squared: 0.1861, Adjusted R-squared: 0.179
F-statistic: 26.29 on 1 and 115 DF, p-value: 1.202e-06

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept)	6.42401750	6.7120054
caratteristiche	0.06023513	0.1360771

Analisi dei residui
Shapiro-Wilk normality test
W = 0.9773, p-value = 0.0448

Trasformazione Boxcox dei dati:

lambda= -0.8

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.2440	2.759e-04	4507.628	< 2e-16	***

Caratteristiche 3.717e-04 7.265e-05 5.116 1.26e-06 ***

Multiple R-squared: 0.1854, Adjusted R-squared: 0.1783
F-statistic: 26.18 on 1 and 115 DF, p-value: 1.261e-06

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept) 1.2430376731 1.2441306208
Caratteristiche 0.0002277994 0.0005156286

Analisi dei residui
Shapiro-Wilk normality test
W = 0.9871, p-value = 0.3287

```
#=====#
#REGRESSIONE TASSE-SUPERFICIE #
#=====#
```

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-28.6666	53.0585	-0.54	0.59	
superficie	0.4973	0.0306	16.25	<2e-16	***

Multiple R-squared: 0.6967, Adjusted R-squared: 0.694
F-statistic: 264.1 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept) -133.765316 76.4321354
superficie 0.436652 0.5578647

Analisi dei residui
Shapiro-Wilk normality test
W = 0.9531, p-value = 0.0004411

Trasformazione logaritmica dei dati:
Regressione log(tasse) ~superficie

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.626e+00	7.190e-02	78.24	<2e-16	***
superficie	5.912e-04	4.146e-05	14.26	<2e-16	***

Multiple R-squared: 0.6387, Adjusted R-squared: 0.6356
F-statistic: 203.3 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept) 5.4834228714 5.7682733924

Analisi dei residui
 Shapiro-Wilk normality test
 W = 0.9198, p-value = 3.036e-06

Trasformazione Boxcox dei dati:

lambda= 0.5

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	33.05383	1.75094	18.88	<2e-16	***
superficie	0.01768	0.00101	17.51	<2e-16	***

Multiple R-squared: 0.7273, Adjusted R-squared: 0.7249
 F-statistic: 306.7 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%
 (Intercept) 29.5855534 36.52210323
 superficie 0.0156815 0.01968154

Analisi dei residui
 Shapiro-Wilk normality test
 W = 0.9178, p-value = 2.335e-06

```
#=====#
#REGRESSIONE TASSE-PREZZO#
#=====#
```

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	67.93981	47.77595	1.422	0.158	
prezzo	0.68294	0.04235	16.128	<2e-16	***

Multiple R-squared: 0.6934, Adjusted R-squared: 0.6908
 F-statistic: 260.1 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%
 (Intercept) -26.6951621 162.5747769

Analisi dei residui
 Shapiro-Wilk normality test
 W = 0.9525, p-value = 0.0004016

Trasformazione logaritmica dei dati:

Regressione log(tasse) ~ prezzo

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.755604	0.066568	86.46	<2e-16	***
prezzo	0.000798	0.000059	13.53	<2e-16	***

Multiple R-squared: 0.614, Adjusted R-squared: 0.6106
F-statistic: 182.9 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%
(Intercept) 5.623746584 5.8874620622
prezzo 0.000681126 0.0009148689

Analisi dei residui
Shapiro-Wilk normality test
W = 0.9524, p-value = 0.0003920

Trasformazione Boxcox dei dati:
lambda=0.6

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	40.32868	3.28312	12.28	<2e-16	***
prezzo	0.04514	0.00291	15.51	<2e-16	***

Multiple R-squared: 0.6767, Adjusted R-squared: 0.6739
F-statistic: 240.7 on 1 and 115 DF, p-value: < 2.2e-16

Intervallo di confidenza per i regressori a livello di significatività del 95%
(Intercept) 33.82545649 46.83191111

Analisi dei residui
Shapiro-Wilk normality test
W = 0.9726, p-value = 0.01689

* REGRESSIONE MULTIPLA * #
#####

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	59.34033	66.34525	0.894	0.373052	
superficie	0.38469	0.06288	6.118	1.49e-08	***
età	-1.32064	1.31273	-1.006	0.316609	
caratteristiche	14.42376	13.55440	1.064	0.289596	
ne1	17.23336	37.56674	0.459	0.647326	
angolo1	-75.70160	44.08776	-1.717	0.088781	.
tasse	0.43006	0.10806	3.980	0.000124	***

Multiple R-squared: 0.7795, Adjusted R-squared: 0.7675
 F-statistic: 64.82 on 6 and 110 DF, p-value: < 2.2e-16

Studio multicollinearità:

Vif

superficie	età	caratteristiche	ne1	angolo1	tasse
3.918505	3.738569	1.193974	1.251001	1.090352	1.031770

Regressione multipla con modello aic:

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	73.13663	56.63965	1.291	0.1992	
superficie	0.37336	0.05945	6.280	6.49e-09	***
angolo1	-82.14183	43.73238	-1.878	0.0629	.
tasse	0.48828	0.09979	4.893	3.32e-06	***

Multiple R-squared: 0.7744, Adjusted R-squared: 0.7684
 F-statistic: 129.3 on 3 and 113 DF, p-value: < 2.2e-16

Analisi dei residui

Shapiro-Wilk normality test

W = 0.937, p-value = 3.409e-05

Intervallo di confidenza per i regressori a livello di significatività del 95%

(Intercept)	-39.0767313	185.3499954
superficie	0.2555736	0.4911486
angolo1	-168.7835566	4.4998917
tasse	0.2905828	0.6859724

Regressione multipla prezzo~ (tasse + superficie)

Coefficienti di regressione:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	62.52216	56.97846	1.097	0.275	
tasse	0.51289	0.10001	5.128	1.21e-06	***
superficie	0.35863	0.05958	6.019	2.18e-08	***

Multiple R-squared: 0.7673, Adjusted R-squared: 0.7633
 F-statistic: 188 on 2 and 114 DF, p-value: < 2.2e-16

Analisi dei residui

Shapiro-Wilk normality test

W = 0.9225, p-value = 4.322e-06

Script

```
#####  
#CARICAMENTO DATASET#  
#####  
  
#caricamento dataset nella variabile hp  
hp=read.table("homeprices.txt",sep="\t",header=T)  
  
#nomi variabili  
names(hp)  
  
#struttura dataset  
str(hp)  
  
#trasformazione tipo variabili categoriali  
hp$ne=factor(hp$ne)  
hp$angolo=factor(hp$angolo)  
str(hp)  
  
#####  
#RICOSTRUZIONE DATI MANCANTI#  
#####  
  
#DATI MANCANTI VARIABILE età  
hp$età  
#posizione dati mancanti  
index=NULL  
#e costruzione vettore dati noti  
dati_noti=NULL  
  
for (i in 1:length(hp$età))  
{  
  if (hp$età[i]=="*")  
    index=c(index,i)  
  else  
    dati_noti=c(dati_noti,as.integer(levels(hp$età))[hp$età[i]])  
}  
index  
str(index)  
dati_noti  
str(dati_noti)  
  
#sampling del vettore  
campione=sample(dati_noti,length(index),replace=TRUE)  
campione  
str(campione)  
  
#conversione in variabile numerica intera  
hp$età=as.integer(levels(hp$età))[hp$età]  
hp$età  
str(hp$età)  
  
#ricostruzione del vettore con l'aggiunta dei nuovi dati  
for (k in 1:length(index))  
{  
  hp$età[index[k]]=campione[k]  
}  
hp$età  
str(hp$età)  
  
str(hp)
```

```

#DATI MANCANTI VARIABILE tasse
hp$tasce
#posizione dati mancanti
index=NULL
#e costruzione vettore dati noti
dati_noti=NULL
for (i in 1:length(hp$tasce))
{
  if (hp$tasce[i]=="*")
    index=c(index,i)
  else
    dati_noti=c(dati_noti,as.integer(levels(hp$tasce))[hp$tasce[i]])
}
str(index)
dati_noti
str(dati_noti)

#sampling del vettore
campione=sample(dati_noti,length(index),replace=TRUE)
str(campione)

#conversione in variabile numerica intera
hp$tasce=as.integer(levels(hp$tasce))[hp$tasce]
hp$tasce
str(hp$tasce)

#ricostruzione del vettore con l'aggiunta dei nuovi dati
for (k in 1:length(index))
{
  hp$tasce[index[k]]=campione[k]
}
hp$tasce
str(hp$tasce)

str(hp)

#attach delle variabili
attach(hp)
str(hp)

#####
#STATISTICA DESCRITTIVA#
#####

#analisi delle singole variabili
summary(hp)

#istogrammi
hist(prezzo,prob=TRUE,main="Prezzo",xlab="x100$",col="azure")
hist(superficie,prob=TRUE,main="Superficie",xlab="square feet",col="pink")
hist(età,prob=TRUE,main="Età",xlab="anni",col="red")
hist(tasse,prob=TRUE,main="Tasse",xlab="x100$",col="blue")

```

```

#boxplot
boxplot(prezzo~angolo,horizontal=TRUE,main="Prezzo vs Angolo",xlab="Prezzo",
ylab="Angolo",col=c('red','green'))
boxplot(prezzo~caratteristiche,horizontal=TRUE,main="Prezzo vs Caratteris-
tiche",xlab="Prezzo",
ylab="Caratteristiche",col=c('blue','yellow','red','green','pink','azure','purple'))
boxplot(prezzo~ne,horizontal=TRUE,main="Prezzo vs Nord-Est",xlab="Prezzo", ylab="Nord-
Est",col=c('yellow','purple'))

boxplot(tasse~angolo,horizontal=TRUE,main="Tasse vs Ango-
lo",xlab="Tasse",ylab="Angolo",col=c('blue','azure'))
boxplot(tasse~caratteristiche,horizontal=TRUE,main="Tasse vs Caratteris-
tiche",xlab="Tasse",ylab="Caratteristiche",col=c('yellow','green','blue','red','pink','az
ure','purple'))
boxplot(tasse~ne,horizontal=TRUE,main="Tasse vs Nord-Est",xlab="Tasse",ylab="Nord-
Est",col=c('yellow','green'))

#####
#REGRESSIONE#
#####

#coeff correlaz
cor(hp)

=====
#REGRESSIONE PREZZO-SUPERFICIE#
=====

plot(superficie,prezzo,ylab="prezzo",xlab="superficie")
regressione_prezzo_superficie=lm(formula = prezzo ~ superficie, data = hp)

summary(regressione_prezzo_superficie)

#retta di regressione
abline(regressione_prezzo_superficie,col="green",lw=2)

#coefficienti di regressione
coef(regressione_prezzo_superficie)

#intervallo di confidenza regressori
confint(regressione_prezzo_superficie,level=0.95)

#analisi residui regressione prezzo-superficie

residuals(regressione_prezzo_superficie)
mean( residuals(regressione_prezzo_superficie))
boxplot(residuals(regressione_prezzo_superficie),horizontal=TRUE,
xlab="Residuals",col=c("red"))
summary(residuals(regressione_prezzo_superficie))

qqnorm(residuals(regressione_prezzo_superficie),ylab="residuals")
qqline(residuals(regressione_prezzo_superficie),col=c("blue"), lwd=2)

hist(residuals(regressione_prezzo_superficie),prob=TRUE,col=c("red"),main="istogramma re-
sidui regr. prezzo-superficie",xlab="Residui")
shapiro.test (residuals(regressione_prezzo_superficie))##pvalue basso non sono normali!!

plot(fitted(regressione_prezzo_superficie),residuals(regressione_prezzo_superficie),xlab=
"Superficie", ylab="Residui")
abline(h=0, col=c("red"),lwd=2)###bisogna normalizzare i dati

```

```

#TRASFORMAZIONE LOGARITMICA

plot(superficie,log(prezzo),xlab="Superficie",ylab="Log prezzo")
regressione_prezzo_superficie_log=lm(formula = log(prezzo)~superficie, data = hp)

summary(regressione_prezzo_superficie_log)

#retta di regressione
abline(regressione_prezzo_superficie_log,col="brown",lwd=2)

#coefficienti di regressione
coef(regressione_prezzo_superficie_log)

#intervallo di confidenza regressori
confint(regressione_prezzo_superficie_log,level=0.95)


#analisi residui regressione prezzo-superficie-logaritmica

residuals(regressione_prezzo_superficie_log)
mean( residuals(regressione_prezzo_superficie_log))
boxplot(residuals(regressione_prezzo_superficie_log),horizontal=TRUE,
xlab="Residuals",col=c("yellow"))
summary(residuals(regressione_prezzo_superficie_log))

qqnorm(residuals(regressione_prezzo_superficie_log),ylab="residuals")
qqline(residuals(regressione_prezzo_superficie_log),col=c("green"), lwd=2)

hist(residuals(regressione_prezzo_superficie_log),prob=TRUE,col=c("purple"),main="istogramma residui regr. log(prezzo)-superficie",xlab="Residui")
shapiro.test( residuals(regressione_prezzo_superficie_log))

plot(fitted(regressione_prezzo_superficie_log),residuals(regressione_prezzo_superficie_log),xlab="Superficie", ylab="Residui")
abline(h=0, col=c("brown"),lwd=2)


#TRASFORMAZIONE BOXCOX

library(MASS)
boxcox(regressione_prezzo_superficie,plotit=T)
bcox=boxcox(regressione_prezzo_superficie,plotit=F)
lambda=bcox$x[which.max(bcox$y)]
lambda

prezzo_trasf=(prezzo^lambda-1)/lambda

plot(superficie,prezzo_trasf)

regressione_prezzo_superficie_modificato=lm(formula = prezzo_trasf ~ superficie, data = hp)

summary(regressione_prezzo_superficie_modificato)

#retta di regressione
abline(regressione_prezzo_superficie_modificato,col="yellow",lwd=2)

#coefficienti di regressione
coef(regressione_prezzo_superficie_modificato)

#intervallo di confidenza regressori
confint(regressione_prezzo_superficie_modificato,level=0.95)

```

```

#analisi residui regressione prezzo-superficie con trasformazione box-cox

residuals(regressione_prezzo_superficie_modificato)
mean( residuals(regressione_prezzo_superficie_modificato))
boxplot(residuals(regressione_prezzo_superficie_modificato),horizontal=TRUE,
xlab="Residuals",col=c("red"))
summary(residuals(regressione_prezzo_superficie_modificato))

qqnorm(residuals(regressione_prezzo_superficie_modificato),ylab="residuals")
qqline(residuals(regressione_prezzo_superficie_modificato),col=c("blue"), lwd=2)
hist(residuals(regressione_prezzo_superficie_modificato),prob=TRUE,col=c("azure"),main="i
stogramma residui regr. prezzo-superficie",xlab="Residui")
shapiro.test (residuals(regressione_prezzo_superficie_modificato))

plot(fitted(regressione_prezzo_superficie_modificato),residuals(regressione_prezzo_superf
icie_modificato),xlab="Superficie", ylab="Residui")
abline(h=0, col=c("red"),lwd=2)


#=====#
#REGRESSIONE PREZZO-CARATTERISTICHE#
#=====#

plot(caratteristiche,prezzo,ylab="prezzo",xlab="caratteristiche")
regressione_prezzo_caratteristiche=lm(formula = prezzo ~ caratteristiche, data = hp)

summary(regressione_prezzo_caratteristiche)

#retta di regressione
abline(regressione_prezzo_caratteristiche,col="red",lw=2)

#coefficienti di regressione
coef(regressione_prezzo_caratteristiche)

#intervallo di confidenza regressori
confint(regressione_prezzo_caratteristiche,level=0.95)

#analisi residui regressione prezzo-caratteristiche

residuals(regressione_prezzo_caratteristiche)
mean( residuals(regressione_prezzo_caratteristiche))
boxplot(residuals(regressione_prezzo_caratteristiche),horizontal=TRUE,
xlab="Residuals",col=c("blue"))
summary(residuals(regressione_prezzo_caratteristiche))

qqnorm(residuals(regressione_prezzo_caratteristiche),ylab="residuals")
qqline(residuals(regressione_prezzo_caratteristiche),col=c("violet"), lwd=2)

hist(residuals(regressione_prezzo_caratteristiche),prob=TRUE,col=c("yellow"),main="istogr
amma residui regr. prezzo-caratteristiche",xlab="Residui")
shapiro.test (residuals(regressione_prezzo_caratteristiche))

plot(fitted(regressione_prezzo_caratteristiche),residuals(regressione_prezzo_caratteri
stiche),xlab="Caratteristiche", ylab="Residui")
abline(h=0, col=c("yellow"),lwd=2)

#TRASFORMAZIONE LOGARITMICA

plot(caratteristiche,log(prezzo),xlab="caratteristiche",ylab="Log prezzo")

```

```

regressione_prezzo_caratteristiche_log=lm(formula = log(prezzo)~caratteristiche, data =
hp)

summary(regressione_prezzo_caratteristiche_log)

#retta di regressione
abline(regressione_prezzo_caratteristiche_log,col="green",lwd=2)

#coefficienti di regressione
coef(regressione_prezzo_caratteristiche_log)

#intervallo di confidenza regressori
confint(regressione_prezzo_caratteristiche_log,level=0.95)

#analisi residui regressione prezzo-caratteristiche-logaritmica

residuals(regressione_prezzo_caratteristiche_log)
mean( residuals(regressione_prezzo_caratteristiche_log))
boxplot(residuals(regressione_prezzo_caratteristiche_log),horizontal=TRUE,
xlab="Residuals",col=c("azure"))
summary(residuals(regressione_prezzo_caratteristiche_log))

qqnorm(residuals(regressione_prezzo_caratteristiche_log),ylab="residuals")
qqline(residuals(regressione_prezzo_caratteristiche_log),col=c("blue"), lwd=2)
hist(residuals(regressione_prezzo_caratteristiche_log),prob=TRUE,col=c("violet"),main="is
togramma residui regr. log(prezzo)-caratteristiche",xlab="Residui")
shapiro.test (residuals(regressione_prezzo_caratteristiche_log))

plot(fitted(regressione_prezzo_caratteristiche_log),residuals(regressione_prezzo_caratter
istiche_log),xlab="Caratteristiche", ylab="Residui")
abline(h=0, col=c("blue"),lwd=2)

#TRASFORMAZIONE BOXCOX
library(MASS)
boxcox(regressione_prezzo_caratteristiche,plotit=T)
bcox=boxcox(regressione_prezzo_caratteristiche,plotit=F)
lambda=bcox$x[which.max(bcox$y)]
lambda

prezzo_trasf=(prezzo^lambda-1)/lambda

plot(caratteristiche,prezzo_trasf)

regressione_prezzo_caratteristiche_modificato=lm(formula = prezzo_trasf ~ caratteristi-
che, data = hp)

summary(regressione_prezzo_caratteristiche_modificato)

#retta di regressione
abline(regressione_prezzo_caratteristiche_modificato,col="brown",lwd=2)

#coefficienti di regressione
coef(regressione_prezzo_caratteristiche_modificato)

#intervallo di confidenza regressori
confint(regressione_prezzo_caratteristiche_modificato,level=0.95)

#analisi residui regressione prezzo-caratteristiche con trasformazione box-cox

residuals(regressione_prezzo_caratteristiche_modificato)

```

```

mean( residuals(regressione_prezzo_caratteristiche_modificato))
boxplot(residuals(regressione_prezzo_caratteristiche_modificato),horizontal=TRUE,
xlab="Residuals",col=c("yellow"))
summary(residuals(regressione_prezzo_caratteristiche_modificato))

qqnorm(residuals(regressione_prezzo_caratteristiche_modificato),ylab="residuals")
qqline(residuals(regressione_prezzo_caratteristiche_modificato),col=c("red"), lwd=2)
hist(residuals(regressione_prezzo_caratteristiche_modificato),prob=TRUE,col=c("azure"),ma
in="istogramma residui regr. prezzo-caratteristiche",xlab="Residui")
shapiro.test (residuals(regressione_prezzo_caratteristiche_modificato))

plot(fitted(regressione_prezzo_caratteristiche_modificato),residuals(regressione_prezzo_c
aratteristiche_modificato),xlab="Caratteristiche", ylab="Residui")
abline(h=0, col=c("purple"),lwd=2)

#=====#
#REGRESSIONE TASSE-SUPERFICIE#
#=====#

plot(superficie,tasse,ylab="Tasse",xlab="Superficie")
regressione_tasse_superficie=lm(formula = tasse ~ superficie, data = hp)

summary(regressione_tasse_superficie)

#retta di regressione
abline(regressione_tasse_superficie,col="purple",lw=2)

#coefficienti di regressione
coef(regressione_tasse_superficie)

#intervallo di confidenza regressori
confint(regressione_tasse_superficie,level=0.95)

#analisi residui regressione tasse-superficie

residuals(regressione_tasse_superficie)
mean( residuals(regressione_tasse_superficie))
boxplot(residuals(regressione_tasse_superficie),horizontal=TRUE,
xlab="Residuals",col=c("red"))
summary(residuals(regressione_tasse_superficie))

qqnorm(residuals(regressione_tasse_superficie),ylab="residuals")
qqline(residuals(regressione_tasse_superficie),col=c("blue"), lwd=2)
hist(residuals(regressione_tasse_superficie),prob=TRUE,col=c("red"),main="istogramma re-
sidui regr. tasse-superficie",xlab="Residui")
shapiro.test (residuals(regressione_tasse_superficie))

plot(fitted(regressione_tasse_superficie),residuals(regressione_tasse_superficie),xlab="S
uperficie", ylab="Residui")
abline(h=0, col=c("red"),lwd=2)

#TRASFORMAZIONE LOGARITMICA

plot(superficie,log(tasse),xlab="Superficie",ylab="Log Tasse")
regressione_tasse_superficie_log=lm(formula = log(tasse)~superficie, data = hp)

summary(regressione_tasse_superficie_log)

#retta di regressione
abline(regressione_tasse_superficie_log,col="brown",lwd=2)

#coefficienti di regressione
coef(regressione_tasse_superficie_log)

```

```

#intervallo di confidenza regressori
confint(regressione_tasse_superficie_log,level=0.95)

#analisi residui regressione tasse-superficie-logaritmica

residuals(regressione_tasse_superficie_log)
mean( residuals(regressione_tasse_superficie_log))
boxplot(residuals(regressione_tasse_superficie_log),horizontal=TRUE,
xlab="Residuals",col=c("purple"))
summary(residuals(regressione_tasse_superficie_log))

qqnorm(residuals(regressione_tasse_superficie_log),ylab="residuals")
qqline(residuals(regressione_tasse_superficie_log),col=c("blue"), lwd=2)
hist(residuals(regressione_tasse_superficie_log),prob=TRUE,col=c("red"),main="istogramma
residui regr. log(tasse)-superficie",xlab="Residui")
shapiro.test (residuals(regressione_tasse_superficie_log))

plot(fitted(regressione_tasse_superficie_log),residuals(regressione_tasse_superficie_log)
,xlab="Superficie", ylab="Residui")
abline(h=0, col=c("red"),lwd=2)

# TRASFORMAZIONE BOXCOX

boxcox(regressione_tasse_superficie,plotit=T)
bcox=boxcox(regressione_tasse_superficie,plotit=F)
lambda=bcox$x[which.max(bcox$y)]
lambda

tasse_trasf=(prezzo^lambda-1)/lambda

plot(superficie,tasse_trasf)

regressione_tasse_superficie_modificato=lm(formula = tasse_trasf ~ superficie, data = hp)

summary(regressione_tasse_superficie_modificato)

#retta di regressione
abline(regressione_tasse_superficie_modificato,col="red",lwd=2)

#coefficienti di regressione
coef(regressione_tasse_superficie_modificato)

#intervallo di confidenza regressori
confint(regressione_tasse_superficie_modificato,level=0.95)

#analisi residui regressione tasse-superficie con trasformazione box-cox

residuals(regressione_tasse_superficie_modificato)
mean( residuals(regressione_tasse_superficie_modificato))
boxplot(residuals(regressione_tasse_superficie_modificato),horizontal=TRUE,
xlab="Residuals",col=c("red"))
summary(residuals(regressione_tasse_superficie_modificato))

qqnorm(residuals(regressione_tasse_superficie_modificato),ylab="residuals")
qqline(residuals(regressione_tasse_superficie_modificato),col=c("blue"), lwd=2)

hist(residuals(regressione_tasse_superficie_modificato),prob=TRUE,col=c("red"),main="isto
gramma residui regr. tasse-superficie",xlab="Residui")
shapiro.test (residuals(regressione_tasse_superficie_modificato))

```

```

plot(fitted(regressione_tasse_superficie_modificato),residuals(regressione_tasse_superficie_modificato),xlab="Superficie", ylab="Residui")
abline(h=0, col=c("green"),lwd=2)

#=====#
#REGRESSIONE TASSE-PREZZO#
#=====#

plot(prezzo,tasse,xlab="Prezzo",ylab="Tasse")
regressione_tasse_prezzo=lm(formula = tasse ~ prezzo, data = hp)

summary(regressione_tasse_prezzo)

#retta di regressione
abline(regressione_tasse_prezzo,col="yellow",lwd=2)

#coefficienti di regressione
coef(regressione_tasse_prezzo)

#intervallo di confidenza regressori
confint(regressione_tasse_prezzo,level=0.95)

#analisi residui regressione tasse-prezzo

residuals(regressione_tasse_prezzo)
mean( residuals(regressione_tasse_prezzo))
boxplot(residuals(regressione_tasse_prezzo),horizontal=TRUE,
xlab="Residuals",col=c("azure"))
summary(residuals(regressione_tasse_prezzo))

qqnorm(residuals(regressione_tasse_prezzo),ylab="residuals")
qqline(residuals(regressione_tasse_prezzo),col=c("green"), lwd=2)
hist(residuals(regressione_tasse_prezzo),prob=TRUE,col=c("green"),main="istogramma residui regr. tasse-prezzo",xlab="Residui")
shapiro.test (residuals(regressione_tasse_prezzo))

plot(fitted(regressione_tasse_prezzo),residuals(regressione_tasse_prezzo),xlab="Prezzo",
ylab="Residui")
abline(h=0, col=c("black"),lwd=2)

# TRASFORMAZIONE LOGARITMICA

plot(prezzo,log(tasse),xlab="Prezzo",ylab="Log tasse")
regressione_tasse_prezzo_log=lm(formula = log(tasse)~prezzo, data = hp)

summary(regressione_tasse_prezzo_log)

#retta di regressione
abline(regressione_tasse_prezzo_log,col="purple",lwd=2)

#coefficienti di regressione
coef(regressione_tasse_prezzo_log)

#intervallo di confidenza regressori
confint(regressione_tasse_prezzo_log,level=0.95)

#analisi residui regressione tasse-prezzo-logaritmica

```

```

residuals(regressione_tasse_prezzo_log)
mean( residuals(regressione_tasse_prezzo_log))
boxplot(residuals(regressione_tasse_prezzo_log),horizontal=TRUE,
xlab="Residuals",col=c("red"))
summary(residuals(regressione_tasse_prezzo_log))

qqnorm(residuals(regressione_tasse_prezzo_log),ylab="residuals")
qqline(residuals(regressione_tasse_prezzo_log),col=c("pink"), lwd=2)
hist(residuals(regressione_tasse_prezzo_log),prob=TRUE,col=c("blue"),main="istogramma re-
sidui regr. log(tasse)-prezzo",xlab="Residui")
shapiro.test (residuals(regressione_tasse_prezzo_log))

plot(fitted(regressione_tasse_prezzo_log),residuals(regressione_tasse_prezzo_log),xlab="P
rezzo", ylab="Residui")
abline(h=0, col=c("red"),lwd=2)

# TRASFORMAZIONE BOXCOX

boxcox(regressione_tasse_prezzo,plotit=T)
bcox=boxcox(regressione_tasse_prezzo,plotit=F)
lambda=bcox$x[which.max(bcox$y)]
lambda

tasse_trasf=(tasse^lambda-1)/lambda

plot(prezzo,tasse_trasf)

regressione_tasse_prezzo_modificato=lm(formula = tasse_trasf ~ prezzo, data = hp)

summary(regressione_tasse_prezzo_modificato)

#retta di regressione
abline(regressione_tasse_prezzo_modificato,col="orange",lwd=2)

#coefficienti di regressione
coef(regressione_tasse_prezzo_modificato)

#intervallo di confidenza regressori
confint(regressione_tasse_prezzo_modificato,level=0.95)

#analisi residui regressione tasse-superficie con trasformazione box-cox

residuals(regressione_tasse_prezzo_modificato)
mean( residuals(regressione_tasse_prezzo_modificato))
boxplot(residuals(regressione_tasse_prezzo_modificato),horizontal=TRUE,
xlab="Residuals",col=c("green"))
summary(residuals(regressione_tasse_prezzo_modificato))

qqnorm(residuals(regressione_tasse_prezzo_modificato),ylab="residuals")
qqline(residuals(regressione_tasse_prezzo_modificato),col=c("blue"), lwd=2)
hist(residuals(regressione_tasse_prezzo_modificato),prob=TRUE,col=c("yellow"),main="istog
ramma residui regr. tasse-prezzo",xlab="Residui")
shapiro.test (residuals(regressione_tasse_prezzo_modificato))

plot(fitted(regressione_tasse_prezzo_modificato),residuals(regressione_tasse_prezzo_modif
icato),xlab="Prezzo", ylab="Residui")
abline(h=0, col=c("orange"),lwd=2)

```

```
#####
# * REGRESSIONE MULTIPLA * #
#####

regressione_multipla=lm(formula = prezzo~.,hp)
summary(regressione_multipla)

#verifica di multicollinearità
library(faraway)
pairs(hp)
multicol<-model.matrix(regressione_multipla)[,-1]
vif(multicol)

modello_aic=step(regressione_multipla)

summary(modello_aic)

# Verifica della normalità dei residui
qqnorm(residuals(modello_aic),ylab="residui")
qqline(residuals(modello_aic),col="blue",lwd=2)
hist(residuals(modello_aic),prob=TRUE,col=c("orange"))
shapiro.test(residuals(modello_aic))

plot(fitted(modello_aic),residuals(modello_aic),xlab="Fitted",ylab="Residui")
abline(h=0,col="violet",lwd=2.5)

#Intervalli di confidenza
confint(modello_aic)

#Regressione multipla prezzo-(tasse e superficie)
regressione_multipla2=lm(formula = prezzo~tasse+superficie,hp)
summary(regressione_multipla2)

# Verifica della normalità dei residui
qqnorm(residuals(regressione_multipla2),ylab="residui")
qqline(residuals(regressione_multipla2),col="green",lwd=2)
hist(residuals(regressione_multipla2),prob=TRUE,col=c("azure"))
shapiro.test(residuals(regressione_multipla2))

plot(fitted(regressione_multipla2),residuals(regressione_multipla2),xlab="Fitted",ylab="Residui")
abline(h=0,col="violet",lwd=2.5)

#####
# * DUMMY * #
#####

plot(tasse,prezzo,col=unclass(angolo))

#regressione prezzo-tasse
posiz_generica=lm(prezzo~tasse,data=hp)

#Regressione sui due subsets
angolo_si=lm(prezzo~tasse,data=hp,subset=(angolo=="1"))
angolo_no=lm(prezzo~tasse,data=hp,subset=(angolo=="0"))
abline(angolo_si,col=1)
abline(angolo_no,col=2)
abline(posiz_generica,col=3)
```