Modelli statistici e statistical learning - Progetto finale

Gruppo PDD

Massimo Ruggiero *264469*
Francesco Romeo *264409*
Giuseppe Zappia *268784*
Domenico Macrì *268798*
Vincenzo Prochilo *264100*
Mauro Schettino *269754*

Indice

1 Introduzione

1.1 Descrizione del dataset

Il dataset in esame, noto come **Housing Data di Boston**, è un insieme di dati che descrive le caratteristiche delle abitazioni nella zona di Boston, Massachusetts. Questo dataset è stato originariamente raccolto e pubblicato da Harrison e Rubinfeld nel 1978 ed è ampiamente utilizzato nella letteratura statistica e nell'apprendimento automatico per scopi didattici e di ricerca. I dati sono stati raccolti attraverso indagini e censimenti, e rappresentano una raccolta di informazioni socio-economiche e ambientali relative alle abitazioni nella zona di Boston.

Il dataset contiene informazioni su 506 osservazioni, ciascuna delle quali rappresenta una zona censuaria (o quartiere) nella zona di Boston. Le variabili includono informazioni sul valore mediano delle case (MEDV), il tasso di criminalità (CRIM), la percentuale di abitazioni occupate dai proprietari (ZN), la concentrazione di ossidi di azoto (NOX), il numero medio di stanze per abitazione (RM), e molte altre variabili che descrivono le caratteristiche delle abitazioni e dell'ambiente circostante.

1.2 Obiettivi del progetto

L'obiettivo principale di questo progetto è quello di analizzare la relazione tra il valore mediano delle case (MEDV) e le altre variabili presenti nel dataset. In particolare, vogliamo costruire un modello di regressione lineare che ci permetta di spiegare come il valore delle case sia influenzato da fattori come il tasso di criminalità, la qualità dell'aria, il numero di stanze, e altre variabili socio-economiche.

La variabile dipendente scelta è **MEDV**, ovvero il valore mediano delle case in migliaia di dollari.

1.3 Risultati attesi

Prima di procedere con l'analisi, ci aspettiamo che alcune variabili abbiano un impatto significativo sul valore delle case. Ad esempio, ci aspettiamo che il numero medio di stanze per abitazione (RM) abbia una relazione positiva con il valore delle case, poiché case più grandi tendono ad avere un valore più alto. Allo stesso modo, ci aspettiamo che il tasso di criminalità (CRIM) abbia un impatto negativo sul valore delle case, poiché aree con un tasso di criminalità più alto tendono ad essere meno desiderabili.

Inoltre, ci aspettiamo che la qualità dell'aria (NOX) e la distanza dai centri di occupazione (DIS) possano influenzare il valore delle case, con una relazione negativa per la concentrazione di ossidi di azoto e una relazione positiva per la distanza dai centri di occupazione.

2 Descrizione del dataset

2.1 Analisi Preliminare

Il dataset non ha necessitato di particolari operazioni di pulizia, perciò ci siamo limitati ad omettere le osservazioni che presentavano valori mancanti, rimuovendole con l'apposita istruzione di R.

2.2 Dimensione campionaria

Il dataset contiene **506 osservazioni** e **14 variabili**. Ogni osservazione rappresenta una zona censuaria nella zona di Boston, e le variabili includono informazioni socio-economiche, ambientali e strutturali relative alle abitazioni.

2.3 Descrizione delle variabili

Le variabili presenti nel dataset sono state raccolte attraverso indagini e censimenti, e rappresentano una vasta gamma di informazioni relative alle abitazioni e all'ambiente circostante. Di seguito è riportata una breve descrizione delle variabili:

Variabile	Descrizione
CRIM	Tasso di criminalità pro capite per città.
ZN	Percentuale di terreni residenziali zonati per lotti di oltre 25.000 piedi quadrati.
INDUS	Percentuale di acri di attività non commerciali per città.
CHAS	Variabile dummy che indica se la zona è confinante con il fiume Charles ($1 = sì$, $0 = no$).
NOX	Concentrazione di ossidi di azoto (parti per 10 milioni).
RM	Numero medio di stanze per abitazione.
AGE	Percentuale di unità abitative occupate dai proprietari costruite prima del 1940.
DIS	Distanza ponderata dai cinque centri di occupazione di Boston.
RAD	Indice di accessibilità alle autostrade radiali.
TAX	Aliquota dell'imposta sulla proprietà per ogni $10.000.
PTRATIO	Rapporto alunni-insegnanti per città.
B	Proporzione di afroamericani per città.
LSTAT	Percentuale di status inferiore della popolazione.
MEDV	Valore mediano delle case in migliaia di dollari (variabile dipendente).

Tabella 1: Descrizione delle variabili nel dataset Housing di Boston.

2.4 Statistiche descrittive

Di seguito sono riportate alcune statistiche descrittive delle variabili presenti nel dataset:

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
CRIM	0.00632	0.08190	0.25372	3.61187	3.56026	88.97620
ZN	0.00	0.00	0.00	11.21	12.50	100.00
INDUS	0.46	5.19	9.69	11.08	18.10	27.74
CHAS	0.00	0.00	0.00	0.06996	0.00	1.00
NOX	0.3850	0.4490	0.5380	0.5547	0.6240	0.8710
RM	3.561	5.886	6.208	6.285	6.623	8.780
AGE	2.90	45.17	76.80	68.52	93.97	100.00
DIS	1.130	2.100	3.207	3.795	5.188	12.127
RAD	1.000	4.000	5.000	9.549	24.000	24.000
TAX	187.0	279.0	330.0	408.2	666.0	711.0
PTRATIO	12.60	17.40	19.05	18.46	20.20	22.00
B	0.32	375.38	391.44	356.67	396.23	396.90
LSTAT	1.730	7.125	11.430	12.715	16.955	37.970
MEDV	5.00	17.02	21.20	22.53	25.00	50.00

Tabella 2: Statistiche descrittive delle variabili nel dataset Housing di Boston.

3 Analisi preliminare

3.1 Distribuzione della variabile dipendente

La variabile dipendente **MEDV** rappresenta il valore mediano delle case in migliaia di dollari. La distribuzione di questa variabile è mostrata nel grafico seguente:

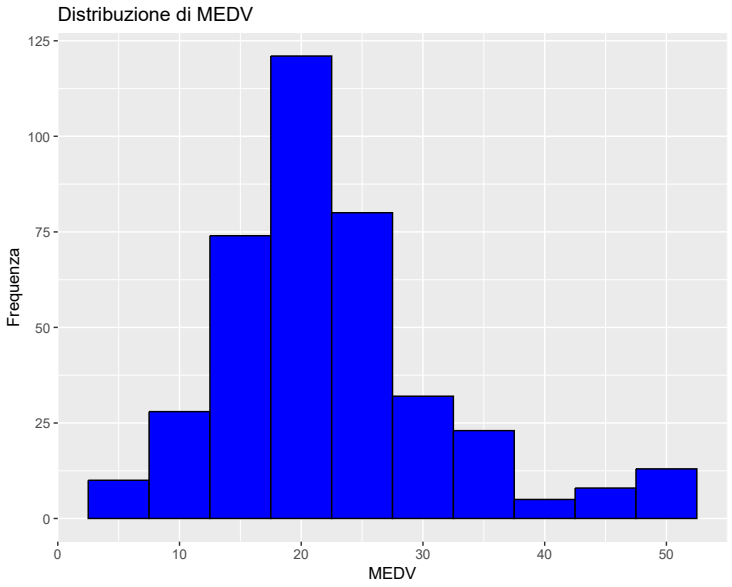

Figura 1: Istogramma della variabile `MEDV`

Dall'istogramma in Figura 1, possiamo osservare che la distribuzione di `MEDV` è approssimativamente normale, con una leggera asimmetria verso destra. La media e la varianza di `MEDV` sono le seguenti:

- Media di `MEDV`: 22.36
- Varianza di `MEDV`: 83.59

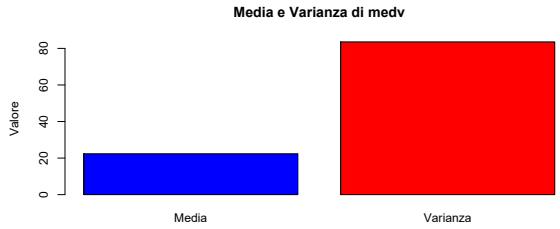

Figura 2: Confronto tra media e varianza di `MEDV`

3.2 Matrice di correlazione

Per comprendere le relazioni tra le variabili indipendenti e la variabile dipendente, è stata calcolata la matrice di correlazione. La matrice di correlazione fornisce una misura della forza e della direzione della relazione lineare tra le variabili. Ci aspettiamo una correlazione positiva tra `RM` (numero medio di stanze per abitazione) e `MEDV`, così come una negativa tra `LSTAT` (percentuale di status basso della popolazione) e la nostra variabile dipendente. Dalla Figura 3, si evince che la

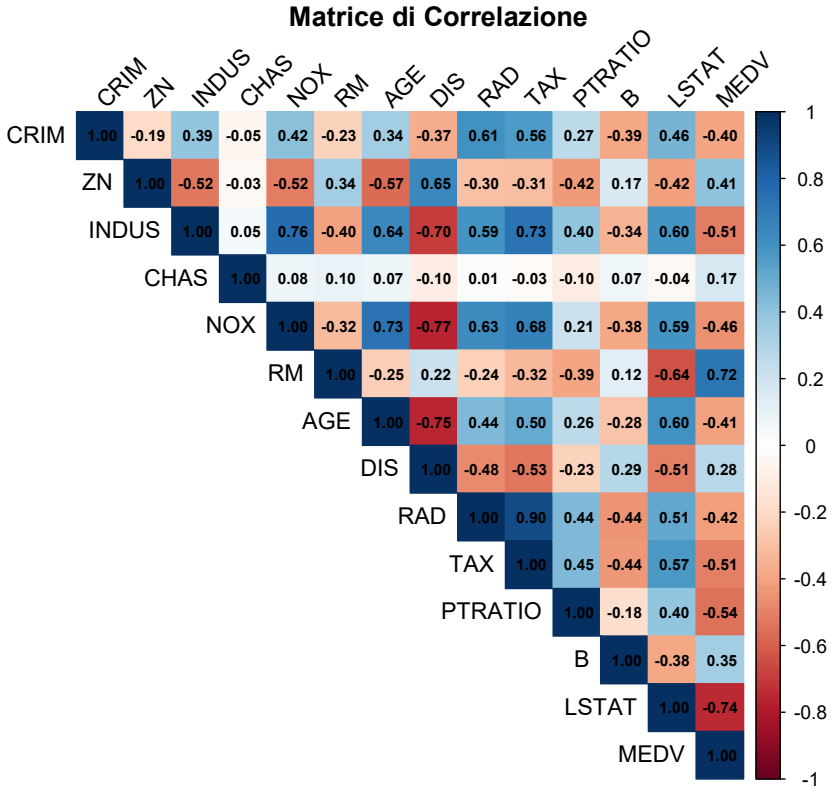

Figura 3: Matrice di correlazione tra le variabili

variabile `RM` ha una forte correlazione positiva con `MEDV`, mentre `LSTAT` ha una forte correlazione negativa. Queste relazioni sono coerenti con le aspettative, poiché un maggior numero di stanze tende ad aumentare il valore della casa, mentre un'alta percentuale di popolazione a basso status tende a diminuirlo.

3.3 Matrice di dispersione

Per visualizzare le relazioni tra le variabili in modo più dettagliato, è stata creata una matrice di dispersione con le rette di regressione.

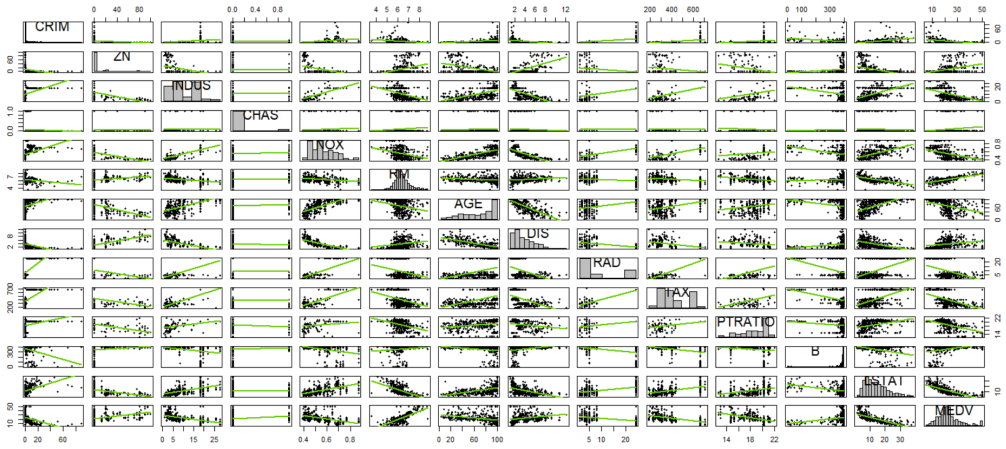

Figura 4: Matrice di dispersione con rette di regressione

La Figura 4 mostra le relazioni tra le variabili, con le rette di regressione che indicano la tendenza lineare. Ad esempio, si può notare che `RM` e `MEDV` hanno una relazione lineare positiva, mentre `LSTAT` e `MEDV` hanno una relazione lineare negativa. Quanto detto finora è confermato effettivamente dal grafico della relazione tra la variabile dipendente e RM ed LSTAT rispettivamente:

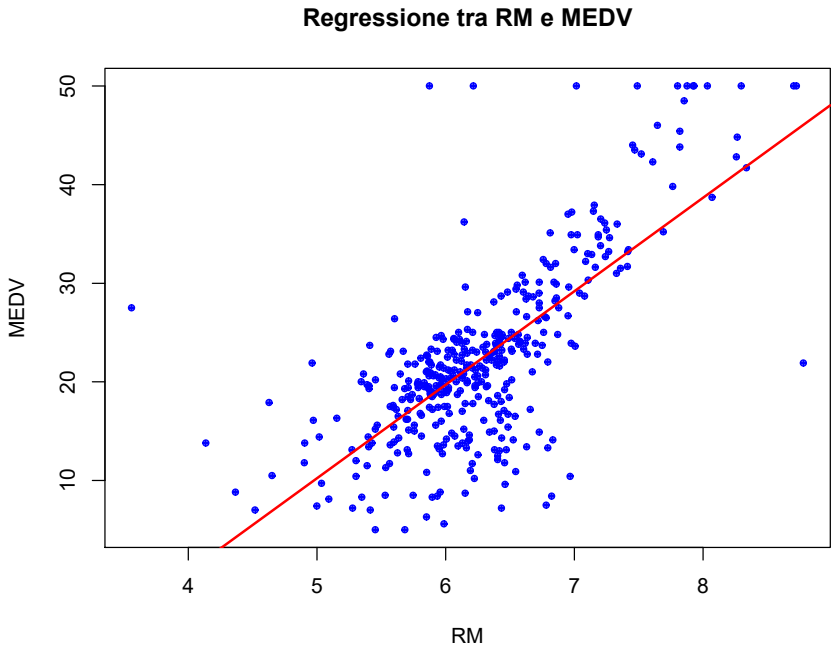

Figura 5: Relazione tra numero medio di stanze per casa e MEDV

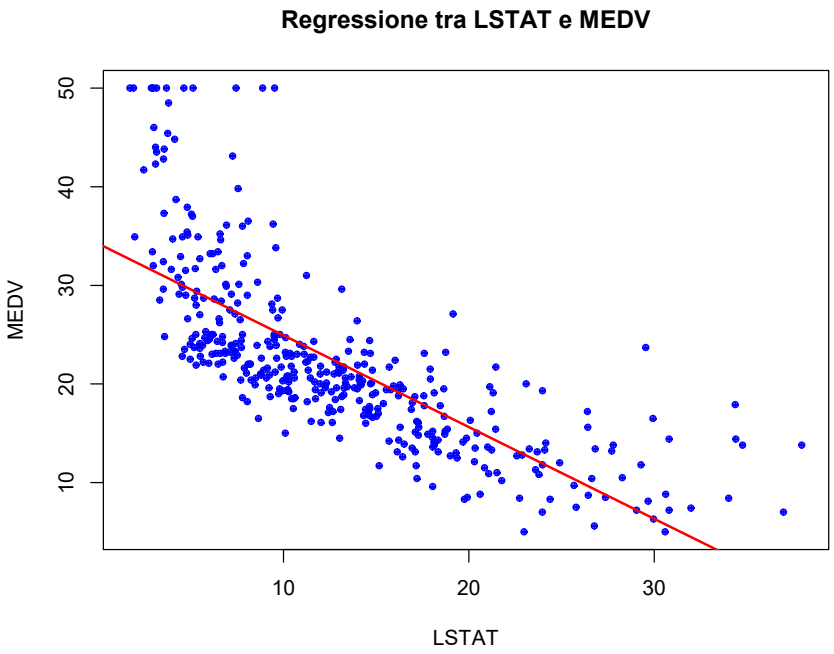

Figura 6: Relazione tra percentuale di status basso della popolazione e MEDV

4 Modelli inferenziali

4.1 Introduzione

L'obiettivo di questa analisi è costruire un modello statistico in grado di spiegare la variabile dipendente `MEDV` (valore medio delle case) in funzione di un insieme di regressori. Il dataset utilizzato contiene informazioni su diverse caratteristiche delle case come già detto.

Il nostro approccio si articola in tre fasi principali:

1. Un primo modello in cui scegliamo manualmente i regressori che riteniamo più rilevanti per spiegare `MEDV`.
2. Un secondo modello in cui includiamo tutti i regressori disponibili per valutare la loro significatività.
3. Un terzo modello finale in cui eliminiamo i regressori non significativi dal modello completo, mantenendo solo quelli che contribuiscono in modo significativo alla spiegazione della variabile dipendente.

4.2 Modello con Regressori scelti da noi (`model1RegressoriSceltiDaNoi`)

```
model1RegressoriSceltiDaNoi <- lm(MEDV ~ (CRIM+RM+DIS+TAX+B+LSTAT), data = housing_data)
summary(model1RegressoriSceltiDaNoi)

#Analisi dei residui
residuiNostroModelloEnormale <- rstandard(model1RegressoriSceltiDaNoi)
hist(residuiNostroModelloEnormale,freq=F,xlim=c(-4,4), ylim=c(0,0.6)); curve(dnorm(x),add=T)

#plottati da soli per capire se ci sono andamenti sistematici negli stessi
windows()
plot(residuiNostroModelloEnormale)

```

```r
SMEDVnostri<-fitted(model1RegressoriSceltiDaNoi)
windows()
Grafico dei valori stimati vs osservati
plot(SMEDVnostri, housing_data$MEDV,
     xlab = "Valori Stimati (SMEDVnostri)",
     ylab = "Valori Osservati (MEDV)",
     main = "Valori Stimati vs Osservati",
     pch = 16, col = "blue")
#Aggiungo la retta a 45 gradi (y = x)
abline(a = 0, b = 1, col = "red", lwd = 2)
```

Il primo modello è stato costruito selezionando manualmente i regressori che riteniamo più rilevanti per spiegare `MEDV`. I regressori scelti sono: `CRIM`, `RM`, `DIS`, `TAX`, `B` e `LSTAT`.

4.2.1 Risultati del Modello

```
Call:
lm(formula = MEDV ~ (CRIM + RM + DIS + TAX + B + LSTAT), data = housing_data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3494  -3.0330  -0.8597   1.7006  28.8543

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.031410   4.245602  -1.185 0.236710
CRIM        -0.045174   0.034891  -1.295 0.196186
RM           5.687836   0.488241  11.650  < 2e-16 ***
DIS         -0.598845   0.151994  -3.940 9.67e-05 ***
TAX         -0.008105   0.002173  -3.729 0.000221 ***
B            0.010239   0.003321   3.083 0.002194 **
LSTAT       -0.490224   0.057645  -8.504 4.03e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.08 on 387 degrees of freedom
Multiple R-squared:  0.696,     Adjusted R-squared:  0.6912
F-statistic: 147.6 on 6 and 387 DF,  p-value: < 2.2e-16
```

Figura 7: Risultati del primo modello

Il modello ha prodotto i seguenti risultati:

- **Statistica F**: 147.6 con un p-value molto basso ($< 2.2e-16$), il che suggerisce che il modello è globalmente significativo e che quindi si può rifiutare l'ipotesi nulla del test F-Fisher, garantendo che almeno uno dei regressori del modello è significativo per spiegare la nostra variabile dipendente.
- **R-quadro (R^2)**: 0.696, il che indica che circa il 69.6% della variabilità della variabile dipendente è spiegata dai regressori inclusi nel modello.
- **R-quadro corretto**: 0.6912, che tiene conto del numero di regressori e fornisce una misura più accurata della bontà di adattamento.

4.2.2 Significatività dei Regressori

- `RM` (numero medio di stanze): coefficiente positivo (5.6878, p-value $< 2e-16$), indicando che un aumento del numero di stanze è associato a un aumento del valore delle case.
- `DIS` (distanza dai centri di occupazione): coefficiente negativo (-0.5988, p-value $9.67e-05$), suggerendo che all'aumentare della distanza, il valore delle case tende a diminuire.
- `TAX` (tassa sulla proprietà): coefficiente negativo (-0.0081, p-value 0.000221), indicando che un aumento delle tasse è associato a una diminuzione del valore delle case.

- `B` (proporzione di afroamericani): coefficiente positivo (0.0102, p-value 0.002194), suggerendo che una maggiore proporzione di afroamericani è associata a un aumento del valore delle case.
- `LSTAT` (percentuale di status basso nella popolazione): coefficiente negativo (-0.4902, p-value $4.03e-16$), indicando che un aumento della percentuale di persone con status basso è associato a una diminuzione del valore delle case.
- `CRIM` (tasso di criminalità): coefficiente negativo (-0.0452, p-value 0.196), non significativo, suggerendo che potrebbe non avere un impatto significativo sul valore delle case in questo modello.

4.2.3 Analisi dei Residui modello con regressori scelti da noi

Si è poi passati all'analisi dei residui per verificare graficamente se le ipotesi fatte in fase di specificazione del modello siano valide oppure no. Inoltre abbiamo plottato l'istogramma dei residui con la normale per capire se l'ipotesi di normalità degli errori è rispettata. L'analisi dei residui mostra che la distribuzione è approssimativamente normale, con un valore medio vicino a zero e una varianza costante. Tuttavia, alcuni residui presentano valori estremi, indicando possibili outlier o una non perfetta aderenza alle ipotesi del modello lineare.

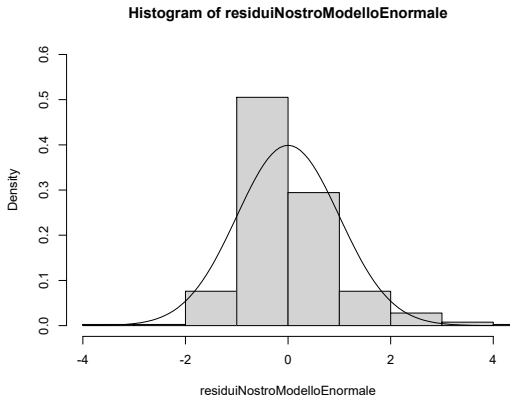

Figura 8: Istogramma dei residui e normale del modello.

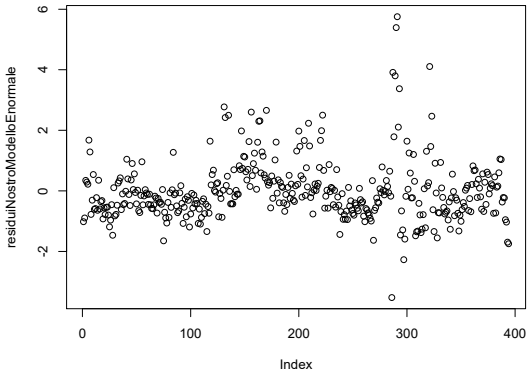

Figura 9: Scatter plot dei residui del modello.

4.2.4 Confronto tra ordinate stimate e osservate

Il seguente grafico riporta il confronto tra le ordinate stimate e quelle osservate per la nostra variabile dipendente. La retta a 45° è importante per capire se il nostro modello rischia di andare incontro ad overfitting, infatti se tutti i punti fossero sulla suddetta retta saremmo proprio in un caso di questi.

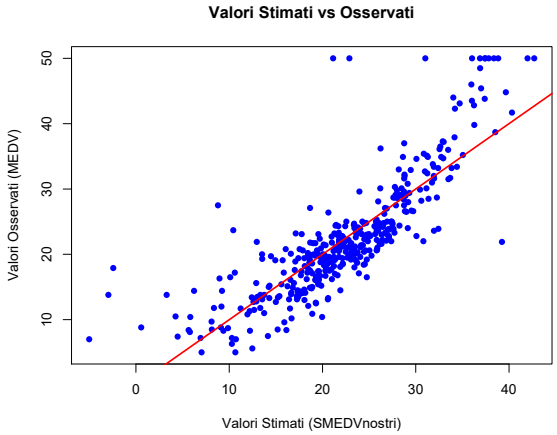

Figura 10: Grafico dei valori stimati ed osservati della variabile dipendente

4.3 Modello con Tutti i Regressori (`model1`)

Il secondo modello include tutti i regressori disponibili nel dataset. Questo approccio ci permette di valutare la significatività di ciascun regressore e di identificare quelli che non contribuiscono in modo significativo alla spiegazione della variabile dipendente.

```
Call:
lm(formula = MEDV ~ ., data = housing_data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4234  -2.5830  -0.5079   1.6681  26.2604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.680059   5.681290   5.752 1.81e-08 ***
CRIM         -0.097594   0.032457  -3.007 0.002815 **
ZN            0.048905   0.014398   3.397 0.000754 ***
INDUS         0.030379   0.065933   0.461 0.645237
CHAS          2.769378   0.925171   2.993 0.002940 **
NOX         -17.969028   4.242856  -4.235 2.87e-05 ***
RM            4.283252   0.470710   9.100  < 2e-16 ***
AGE          -0.012991   0.014459  -0.898 0.369504
DIS          -1.458510   0.211007  -6.912 2.03e-11 ***
RAD           0.285866   0.069298   4.125 4.55e-05 ***
TAX          -0.013146   0.003955  -3.324 0.000975 ***
PTRATIO      -0.914582   0.140581  -6.506 2.44e-10 ***
B             0.009656   0.002970   3.251 0.001251 **
LSTAT        -0.423661   0.055022  -7.700 1.19e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.487 on 380 degrees of freedom
Multiple R-squared:  0.7671,    Adjusted R-squared:  0.7591
F-statistic: 96.29 on 13 and 380 DF,  p-value: < 2.2e-16
```

Figura 11: Risultati modello con tutti i regressori

4.3.1 Risultati del Modello

- **Statistica F**: 96.29 con un p-value molto basso ($< 2.2e-16$), confermando la significatività globale del modello.

- **R-quadro (R^2)**: 0.7671, leggermente superiore al modello precedente, indicando che l'inclusione di tutti i regressori aumenta la capacità esplicativa del modello.
- **R-quadro corretto**: 0.7591, che tiene conto del numero di regressori e conferma che il modello è ancora ben adattato.

4.3.2 Analisi dei Residui modello con tutti i regressori

Anche in questo caso abbiamo visualizzato i residui di questo modello:

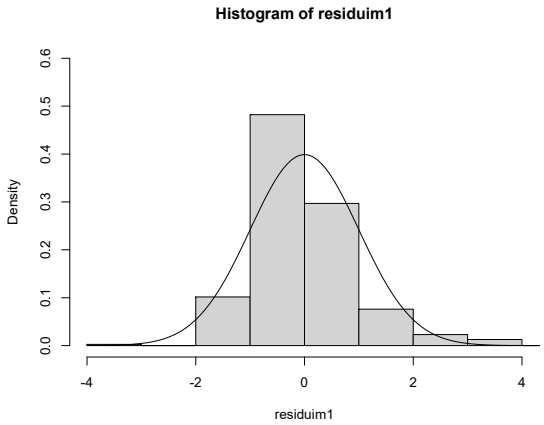

Figura 12: Istogramma dei residui e normale del modello.

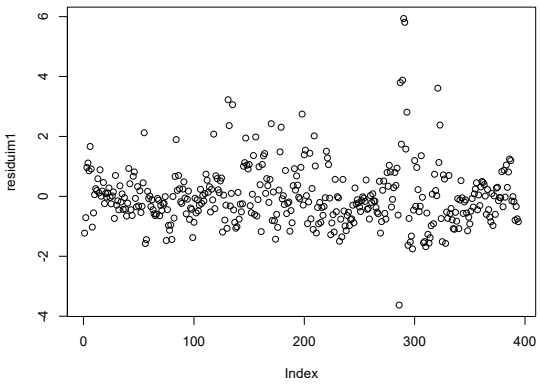

Figura 13: Scatter plot dei residui del modello.

4.3.3 Significatività dei Regressori

Tra i regressori inclusi, i seguenti risultano statisticamente significativi:

- `CRIM`: coefficiente negativo (-0.0976, p-value 0.002815), indicando che un aumento del tasso di criminalità è associato a una diminuzione del valore delle case.
- `RM`: coefficiente positivo (4.2833, p-value $< 2e-16$), confermando che un aumento del numero di stanze aumenta il valore delle case.
- `DIS`: coefficiente negativo (-1.4585, p-value $2.03e-11$), suggerendo che la distanza dai centri di occupazione ha un impatto negativo sul valore delle case.
- `TAX`: coefficiente negativo (-0.0131, p-value 0.000975), indicando che tasse più elevate sono associate a valori delle case più bassi.
- `LSTAT`: coefficiente negativo (-0.4237, p-value $1.19e-13$), confermando che una maggiore percentuale di status basso nella popolazione riduce il valore delle case.

I regressori `INDUS` e `AGE` non risultano significativi (p-value > 0.05), suggerendo che potrebbero essere eliminati dal modello.

4.4 Modello senza i regressori non significativi

Il modello finale è stato ottenuto eliminando proprio i regressori non significativi dal modello completo. Questo approccio ci permette di mantenere solo i regressori che contribuiscono in modo significativo alla spiegazione della variabile dipendente, verificando se si ha un miglioramento nel modello.

```
Call:
lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
    TAX + PTRATIO + B + LSTAT, data = housing_data)

Residuals:
    Min      1Q  Median      3Q     Max
-15.214  -2.552  -0.503   1.768  26.027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.975051   5.630782   5.856 1.02e-08 ***
CRIM         -0.098151   0.032405  -3.029 0.002621 **
ZN            0.049962   0.014169   3.526 0.000473 ***
CHAS          2.788061   0.919721   3.031 0.002600 **
NOX         -18.467815   3.895303  -4.741 3.01e-06 ***
RM            4.166982   0.455473   9.149  < 2e-16 ***
DIS          -1.420599   0.197272  -7.201 3.20e-12 ***
RAD           0.282322   0.065525   4.309 2.09e-05 ***
TAX          -0.012400   0.003471  -3.573 0.000398 ***
PTRATIO      -0.914756   0.138631  -6.599 1.39e-10 ***
B             0.009477   0.002961   3.201 0.001483 **
LSTAT        -0.439994   0.051567  -8.532 3.41e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.481 on 382 degrees of freedom
Multiple R-squared:  0.7665,    Adjusted R-squared:  0.7598
F-statistic:   114 on 11 and 382 DF,  p-value: < 2.2e-16
```

Figura 14: Risultati ultimo modello stimato

4.4.1 Risultati del Modello

- **Statistica F**: 114 con un p-value molto basso ($< 2.2e-16$), confermando la significatività globale del modello.
- **R-quadro (R^2)**: 0.7665, leggermente inferiore al modello completo, ma con un R-quadro corretto più alto (0.7598), indicando un migliore adattamento.

4.4.2 Interpretazione dei Regressori

Nel modello finale, i regressori sono tutti significativi.

4.5 Conclusioni

Il modello finale, `model2`, rappresenta il miglior compromesso tra complessità e capacità esplicativa. I regressori inclusi nel modello sono tutti statisticamente significativi e forniscono una buona spiegazione della variabilità della variabile dipendente `MEDV`. L'interpretazione dei coefficienti ci permette di comprendere come ciascun regressore impatti sul valore delle case, fornendo indicazioni utili per decisioni politiche o economiche relative al mercato immobiliare.

4.6 Best Subset Selection

Oltre ai modelli costruiti manualmente, abbiamo utilizzato la tecnica di **Best Subset Selection** per identificare automaticamente il miglior insieme di regressori in grado di spiegare la variabile dipendente `MEDV`. Questo approccio ci permette di confrontare i risultati ottenuti con la selezione manuale dei regressori e di valutare se esiste un sottoinsieme ottimale di variabili che migliora ulteriormente la capacità predittiva del modello.

```r
best_subset <- ols_step_best_subset(model2)
print(best_subset)
plot(best_subset)
```

Di seguito vengono riportati i risultati ottenuti:

Model Index	Predictors
1	LSTAT
2	RM, LSTAT
3	RM, PTRATIO, LSTAT
4	RM, PTRATIO, B, LSTAT
5	NOX, RM, DIS, PTRATIO, LSTAT
6	CHAS, NOX, RM, DIS, PTRATIO, LSTAT
7	CHAS, NOX, RM, DIS, PTRATIO, B, LSTAT
8	ZN, CHAS, NOX, RM, DIS, PTRATIO, B, LSTAT
9	CRIM, ZN, CHAS, NOX, RM, DIS, PTRATIO, B, LSTAT
10	ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B, LSTAT
11	CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B, LSTAT

Tabella 3: Best Subsets Regression

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.5527	0.5516	0.5466	341.7434	2549.9570	1429.5612	2561.8861	14769.3113	37.6758	0.0959	0.4518
2	0.6585	0.6568	0.6480	170.6535	2445.6163	1325.5506	2461.5217	11304.6238	28.9102	0.0736	0.3467
3	0.7037	0.7014	0.6932	98.7994	2391.7494	1272.0573	2411.6311	9835.3701	25.2160	0.0642	0.3024
4	0.7157	0.7128	0.7037	81.1278	2377.4286	1257.7291	2401.2868	9460.6011	24.3159	0.0619	0.2916
5	0.7312	0.7277	0.7174	57.7442	2357.3178	1237.9243	2385.1522	8967.4165	23.1059	0.0588	0.2771
6	0.7405	0.7365	0.7231	44.5602	2345.4729	1226.3312	2377.2837	8680.2091	22.4217	0.0571	0.2689
7	0.7472	0.7426	0.7283	35.6384	2337.2046	1218.3169	2372.9918	8478.8775	21.9561	0.0559	0.2633
8	0.7524	0.7473	0.7324	29.0682	2330.9555	1212.3386	2370.7190	8324.8247	21.6106	0.0550	0.2592
9	0.7550	0.7493	0.7325	26.8222	2328.8037	1210.3392	2372.5436	8259.0706	21.4930	0.0547	0.2578
10	0.7609	0.7546	0.7389	19.1743	2321.2040	1203.2393	2368.9202	8081.3720	21.0825	0.0537	0.2528
11	0.7665	0.7598	0.7432	12.0000	2313.8534	1196.4818	2365.5460	7912.5521	20.6929	0.0527	0.2482

Tabella 4: Subsets Regression Summary

Come si può notare, il modello con tutti i regressori risulta essere quello che ha le migliori prestazioni, infatti massimizza l'R-quadro corretto e minimizza anche l'AIC.

4.7 Multicollinearità

La multicollinearità è un fenomeno che si verifica quando due o più variabili indipendenti in un modello di regressione sono altamente correlate tra loro. Questo può causare problemi nella stima dei coefficienti di regressione, rendendoli instabili e difficili da interpretare. In particolare, la multicollinearità può portare a:

- **Instabilità dei coefficienti**: Piccole variazioni nei dati possono causare grandi cambiamenti nei coefficienti stimati.
- **Inflazione degli errori standard**: Gli errori standard dei coefficienti diventano più grandi, rendendo difficile determinare la significatività statistica delle variabili.
- **Difficoltà nell'interpretazione**: Quando due variabili sono altamente correlate, diventa difficile distinguere l'effetto individuale di ciascuna variabile sulla variabile dipendente.

Per identificare la presenza di multicollinearità, è stato calcolato il **Variance Inflation Factor (VIF)** per ciascuna variabile indipendente. Il VIF misura quanto la varianza di un coefficiente di regressione è inflata a causa della correlazione con altre variabili indipendenti. La formula per il VIF è:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

dove R_j^2 è il coefficiente di determinazione ottenuto regredendo la variabile X_j su tutte le altre variabili indipendenti. Un VIF elevato indica che la variabile X_j è altamente correlata con le altre variabili indipendenti.

4.7.1 Interpretazione dei Risultati del VIF

I valori di VIF ottenuti per le variabili del dataset sono riportati di seguito:

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
1.7414	2.3218	4.0497	1.0692	4.4958	2.1070	3.1738	3.8274	6.9867	8.6514	1.8106	1.3723	3.1563

Tabella 5: VIF associati ai regressori

4.7.2 Commento sui Risultati

- **Valori di VIF inferiori a 5**: La maggior parte delle variabili ha un VIF inferiore a 5, il che indica che non vi è una multicollinearità significativa. Ad esempio, `CRIM` (1.74), `ZN` (2.32), `CHAS` (1.07), `RM` (2.11), `PTRATIO` (1.81) e `B` (1.37) hanno valori di VIF molto bassi, suggerendo che queste variabili sono poco correlate con le altre variabili indipendenti.
- **Valori di VIF tra 5 e 10**: Alcune variabili, come `INDUS` (4.05), `NOX` (4.50), `AGE` (3.17), `DIS` (3.83), `RAD` (6.99) e `LSTAT` (3.16), presentano valori di VIF leggermente più elevati, ma comunque al di sotto della soglia di 10. Questo suggerisce una moderata correlazione con altre variabili, ma non sufficiente per causare problemi significativi di multicollinearità.
- **Valori di VIF superiori a 10**: La variabile `TAX` ha un VIF di 8.65, che è vicino alla soglia di 10. Questo indica che `TAX` è moderatamente correlata con altre variabili indipendenti, ma non al punto da rendere il modello instabile. Tuttavia, è importante monitorare questa variabile durante la costruzione del modello per evitare potenziali problemi.

4.7.3 Conclusioni sull'Analisi del VIF

In generale, i valori di VIF ottenuti indicano che non vi è un problema significativo di multicollinearità nel dataset. La maggior parte delle variabili ha un VIF inferiore a 5, e solo alcune variabili presentano valori leggermente più elevati, ma comunque al di sotto della soglia critica di 10. Questo suggerisce che il modello di regressione può essere stimato in modo affidabile senza dover rimuovere variabili a causa della multicollinearità.

Tuttavia, è consigliabile prestare attenzione alle variabili con VIF più elevati (ad esempio, `TAX` e `RAD`) durante la fase di selezione del modello, poiché potrebbero influenzare la stabilità dei coefficienti.

5 Eteroschedasticità

5.1 Introduzione all'Eteroschedasticità

L'eteroschedasticità è un fenomeno che si verifica quando la varianza degli errori in un modello di regressione non è costante, ma varia al variare delle variabili indipendenti. Questo può portare a stime dei coefficienti non efficienti e a errori standard distorti, compromettendo l'affidabilità dei test statistici. Per verificare la presenza di eteroschedasticità, sono stati eseguiti diversi test, tra cui il test di Breusch-Pagan e il test di White.

5.2 Test di Breusch-Pagan

Il test di Breusch-Pagan è un test statistico utilizzato per rilevare la presenza di eteroschedasticità. Il test si basa sulla regressione dei residui al quadrato del modello originale rispetto alle variabili indipendenti. Se il modello risultante è significativo, si conclude che c'è eteroschedasticità.

```
1
2 #Calcolo dei residui al quadrato del modello
3 residui_m2 <- resid(model1)
4 residui_m2_quad <- residui_m2^2
5
```

```r
#Creazione di un modello ausiliario in cui i residui al quadrato #sono la
    variabile dipendente
#e TUTTI i regressori originali sono le variabili indipendenti
modello_ausiliario_BP <- lm(residui_m2_quad ~ ., data = housing_data)

#Esecuzione del test di Breusch-Pagan
summary(modello_ausiliario_BP)
```

5.2.1 Risultati del Test di Breusch-Pagan

Nel nostro caso, il test di Breusch-Pagan è stato eseguito sui residui del modello iniziale. I risultati del test sono i seguenti:

```
Call:
lm(formula = residui_m2_quad ~ ., data = housing_data)

Residuals:
   Min     1Q Median     3Q    Max
-78.53 -20.53  -5.52  12.33 433.49

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -116.18804   57.94984  -2.005 0.045676 *
CRIM           0.68669    0.32129   2.137 0.033212 *
ZN            -0.17541    0.14298  -1.227 0.220659
INDUS         -0.06213    0.64521  -0.096 0.923333
CHAS          14.56890    9.15711   1.591 0.112445
NOX           35.40757   42.47653   0.834 0.405043
RM           -36.04918    5.08200  -7.093 6.46e-12 ***
AGE            0.36307    0.14160   2.564 0.010732 *
DIS            5.63689    2.19023   2.574 0.010442 *
RAD           -0.58102    0.69296  -0.838 0.402295
TAX            0.13095    0.03925   3.336 0.000934 ***
PTRATIO        4.79141    1.44989   3.305 0.001041 **
B             -0.05791    0.02945  -1.966 0.050026 .
LSTAT          1.25095    0.57875   2.161 0.031285 *
MEDV           7.30328    0.50186  14.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.9 on 379 degrees of freedom
Multiple R-squared:  0.4312,    Adjusted R-squared:  0.4102
F-statistic: 20.53 on 14 and 379 DF,  p-value: < 2.2e-16
```

Il p-value del test è inferiore a 0.05, il che ci porta a rifiutare l'ipotesi nulla di omoschedasticità. Pertanto, possiamo concludere che c'è eteroschedasticità nel modello.

5.3 Test di White

Per avere un'ulteriore conferma del risultato ottenuto, effettuiamo anche il test di White che è un altro metodo per rilevare l'eteroschedasticità. Questo test si basa sulla regressione dei residui al quadrato rispetto ai valori stimati e ai loro quadrati. Se il modello risultante è significativo, si conclude che c'è eteroschedasticità.

```r
#Calcolo delle ordinate stimate (fitted values) e dei loro quadrati
fitted_m2 <- fitted(model1)
fitted_m2_quad <- fitted_m2^2

```

```r
#Creazione di un modello ausiliario per il test di White
#I residui al quadrato sono la variabile dipendente, mentre le #ordinate stimate e i loro quadrati sono le variabili indipendenti
modello_ausiliario_White <- lm(residui_m2_quad ~ fitted_m2 + fitted_m2_quad)

#Esecuzione del test di White
summary(modello_ausiliario_White)
```

5.3.1 Risultati del Test di White

Il test di White è stato eseguito sui residui del modello iniziale. I risultati del test sono i seguenti:

```
Call:
lm(formula = residui_m2_quad ~ fitted_m2 + fitted_m2_quad)

Residuals:
   Min     1Q Median     3Q    Max
-62.20 -13.58  -9.73  -2.12 677.07

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    63.33473   16.64875   3.804 0.000165 ***
fitted_m2      -4.95502    1.46167  -3.390 0.000770 ***
fitted_m2_quad  0.11859    0.03074   3.858 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.1 on 391 degrees of freedom
Multiple R-squared:  0.04173,   Adjusted R-squared:  0.03683
F-statistic: 8.513 on 2 and 391 DF,  p-value: 0.0002404
```

Anche in questo caso, il p-value è inferiore a 0.05, confermando la presenza di eteroschedasticità.

5.4 Trasformazione Logaritmica della Variabile Dipendente

Per affrontare il problema dell'eteroschedasticità, è stata applicata una trasformazione logaritmica alla variabile dipendente *MEDV*. Successivamente, il modello è stato ristimato e il test di Breusch-Pagan è stato ripetuto.

5.4.1 Risultati del Test di Breusch-Pagan dopo la Trasformazione

I risultati del test di Breusch-Pagan dopo la trasformazione logaritmica sono i seguenti:

```
Call:
lm(formula = residui_m1_log_quad ~ ., data = housing_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.061541 -0.004486 -0.000031  0.003024  0.103454

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.330e-01  3.087e-02   7.547 3.35e-13 ***
CRIM         5.080e-04  1.182e-04   4.298 2.20e-05 ***
ZN          -4.391e-05  4.752e-05  -0.924 0.356155
INDUS        2.835e-04  2.138e-04   1.326 0.185716
```

```
CHAS         3.785e-03  3.033e-03   1.248 0.212857
NOX         -3.431e-02  1.413e-02  -2.428 0.015641 *
RM          -6.790e-04  1.743e-03  -0.389 0.697134
AGE         -5.241e-05  4.691e-05  -1.117 0.264511
DIS         -5.051e-04  7.248e-04  -0.697 0.486268
RAD         -4.202e-04  2.308e-04  -1.821 0.069408 .
TAX          2.628e-05  1.307e-05   2.010 0.045172 *
PTRATIO     -5.235e-04  4.805e-04  -1.089 0.276645
B            3.660e-05  9.765e-06   3.748 0.000206 ***
LSTAT       -1.288e-05  2.152e-04  -0.060 0.952278
MEDV         3.769e-03  3.463e-04  10.884  < 2e-16 ***
log_MEDV    -9.773e-02  8.428e-03 -11.595  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01452 on 378 degrees of freedom
Multiple R-squared:  0.5259,    Adjusted R-squared:  0.5071
F-statistic: 27.96 on 15 and 378 DF,  p-value: < 2.2e-16
```

Il p-value è ancora inferiore a 0.05, indicando che la trasformazione logaritmica non ha risolto completamente il problema dell'eteroschedasticità.

5.5 Utilizzo di Errori Robusti

Per correggere l'eteroschedasticità, sono stati utilizzati errori robusti, che forniscono stime più affidabili dei coefficienti e degli errori standard. I coefficienti con errori robusti sono i seguenti:

```
Coefficienti con errori robusti:
               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  32.6800585   8.6299007  3.7868 0.0001773 ***
CRIM         -0.0975938   0.0314289 -3.1052 0.0020440 **
ZN            0.0489049   0.0148691  3.2890 0.0010991 **
INDUS         0.0303790   0.0568419  0.5344 0.5933448
CHAS          2.7693781   1.5043057  1.8410 0.0664057 .
NOX         -17.9690282   4.0545335 -4.4318 1.224e-05 ***
RM            4.2832519   0.9207643  4.6518 4.549e-06 ***
AGE          -0.0129908   0.0163256 -0.7957 0.4266861
DIS          -1.4585100   0.2350380 -6.2054 1.428e-09 ***
RAD           0.2858656   0.0636008  4.4947 9.263e-06 ***
TAX          -0.0131464   0.0030580 -4.2991 2.183e-05 ***
PTRATIO      -0.9145824   0.1300381 -7.0332 9.445e-12 ***
B             0.0096557   0.0030073  3.2107 0.0014367 **
LSTAT        -0.4236607   0.1003734 -4.2208 3.048e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L'utilizzo di errori robusti ha permesso di ottenere stime più affidabili dei coefficienti, anche in presenza di eteroschedasticità, in modo da poter usare il nostro modello anche per test e intervalli di confidenza.

6 Modelli predittivi

In questa sezione metteremo a confronto alcuni modelli di regolarizzazione ottenuti a partire dai nostri dati. Le tecniche di regolarizzazione sono fondamentali per ridurre l'overfitting, specialmente quando si ha a che fare con un numero elevato di regressori o quando i regressori sono fortemente correlati. Le tecniche utilizzate sono: **Ridge Regression**, **LASSO** e **Elastic Net**.

6.1 Teoria delle Tecniche di Regolarizzazione

Le tecniche di regolarizzazione introducono una penalizzazione nella funzione di costo del modello, con l'obiettivo di ridurre la complessità del modello e migliorare la sua capacità di generalizzazione. La funzione di costo penalizzata è definita come:

$$PLS(\beta) = (y - X\beta)^T(y - X\beta) + \lambda \cdot \text{pen}(\beta)$$

dove:

- y è il vettore delle osservazioni della variabile dipendente,
- X è la matrice dei regressori,
- β è il vettore dei coefficienti di regressione,
- λ è il parametro di penalizzazione,
- $\text{pen}(\beta)$ è il termine di penalizzazione, che varia a seconda della tecnica utilizzata.

6.1.1 Ridge Regression

La **Ridge Regression** utilizza una penalizzazione L_2, ovvero la somma dei quadrati dei coefficienti:

$$\text{pen}(\beta) = \sum_{j=1}^{k} \beta_j^2$$

Questa penalizzazione riduce l'entità dei coefficienti senza eliminarli completamente, rendendo il modello più stabile in presenza di multicollinearità.

6.1.2 LASSO

Il **LASSO** (Least Absolute Shrinkage and Selection Operator) utilizza una penalizzazione L_1, ovvero la somma dei valori assoluti dei coefficienti:

$$\text{pen}(\beta) = \sum_{j=1}^{k} |\beta_j|$$

Questa penalizzazione non solo riduce l'entità dei coefficienti, ma può anche portare all'esclusione di alcuni regressori dal modello, effettuando una selezione delle variabili.

6.1.3 Elastic Net

L'**Elastic Net** combina le penalizzazioni L_1 e L_2 di LASSO e Ridge, bilanciando tra la selezione delle variabili e la riduzione dei coefficienti. La penalizzazione è definita come:

$$\text{pen}(\beta) = \alpha \sum_{j=1}^{k} |\beta_j| + (1 - \alpha) \sum_{j=1}^{k} \beta_j^2$$

dove α è un parametro che controlla il bilanciamento tra le due penalizzazioni. Per $\alpha = 0$, Elastic Net si riduce a Ridge Regression, mentre per $\alpha = 1$ si riduce a LASSO.

6.2 Preparazione dei Dati

Prima di applicare le tecniche di regolarizzazione, i dati sono stati preparati come segue:

- Le variabili indipendenti X sono state separate dalla variabile dipendente y (MEDV, il prezzo delle case).
- I regressori sono stati standardizzati per garantire che tutte le variabili abbiano la stessa scala, una pratica comune nella regolarizzazione.
- È stata creata una griglia di valori per il parametro di penalizzazione λ, che varia da 10^{10} a 10^{-2}.

6.3 Risultati delle Tecniche di Regolarizzazione

6.3.1 Cross-Validation

Per la Ridge Regression, la Lasso e la Elastic Net, è stata eseguita una cross-validation a 10 fold per determinare il valore ottimale di λ.

6.3.2 Confronto dei Modelli

Per confrontare le prestazioni dei tre modelli, è stato calcolato l'MSE (Mean Squared Error) i cui risultati sono riassunti nella seguente tabella:

Modello	**MSE**
Ridge Regression	5.287105
LASSO	5.341359
Elastic Net	5.040578

Tabella 6: Confronto dei modelli in termini di MSE.

Il modello migliore è stato selezionato in base al valore minimo di MSE. In questo caso, **Elastic Net** ha ottenuto il minor MSE (**5.040578**), seguito da Ridge Regression e LASSO . Questo suggerisce che Elastic Net è il modello più adatto per prevedere il prezzo delle case (MEDV) in questo dataset, grazie alla sua capacità di bilanciare la selezione delle variabili e la riduzione dei coefficienti. Le tecniche di regolarizzazione hanno dimostrato di essere efficaci nel migliorare le prestazioni del modello di regressione lineare. In particolare, Elastic Net ha fornito il miglior compromesso tra selezione delle variabili e riduzione dei coefficienti, ottenendo il minor errore di previsione. Questo risultato è in linea con le aspettative, dato che Elastic Net combina i vantaggi di Ridge e LASSO, rendendolo particolarmente adatto per casistiche come quella del nostro dataset.

6.4 Confronto dei Regressori

Un aspetto cruciale delle tecniche di regolarizzazione è il loro effetto sui coefficienti dei regressori. Di seguito è riportata la tabella che specifica l'apporto ai coefficienti del dataset di ciascuna di queste tecniche.

Variable	Ridge	LASSO	ElasticNet
(Intercept)	22.3596	22.3596	22.3596
CRIM	0.9693	0.9162	0.9295
ZN	0.5542	0.4730	0.5265
INDUS	-0.1584	0.0000	-0.0463
CHAS	0.0695	0.0220	0.0518
NOX	-0.0711	0.0000	0.0000
RM	1.5115	1.3935	1.4539
AGE	-0.2428	0.0000	-0.0876
DIS	-0.8118	-0.5174	-0.6294
RAD	0.0365	0.0000	0.0126
TAX	0.0896	0.0000	0.0000
PTRATIO	-0.3289	-0.2955	-0.3155
B	0.0656	0.0000	0.0021
LSTAT	1.0288	0.6239	0.7504
log_MEDV	8.7476	8.5649	8.5806

Tabella 7: Regressione Ridge, LASSO ed ElasticNet

6.4.1 Ridge Regression

Nella Ridge Regression, tutti i regressori vengono mantenuti nel modello, ma i loro coefficienti vengono ridotti. Questo è evidente dai risultati ottenuti, nessun regressore è stato escluso, ma i coefficienti sono stati ridotti in modo significativo rispetto al modello di regressione lineare non penalizzato.

6.4.2 LASSO

Il LASSO, invece, ha escluso diversi regressori dal modello, impostando i loro coefficienti a zero mentre i regressori mantenuti hanno coefficienti ridotti. Questo dimostra la capacità del LASSO di effettuare una selezione delle variabili, escludendo i regressori meno rilevanti.

6.4.3 Elastic Net

L'Elastic Net combina le caratteristiche di Ridge e LASSO, riducendo i coefficienti e, in alcuni casi, escludendo i regressori. Elastic Net ha mantenuto un numero maggiore di regressori rispetto al LASSO, ma ha comunque escluso quelli meno rilevanti, dimostrando un buon bilanciamento tra selezione delle variabili e riduzione dei coefficienti.

6.5 Visualizzazione grafica dei risultati

I risultati ottenuti possono essere visualizzati graficamente grazie ai seguenti plot:

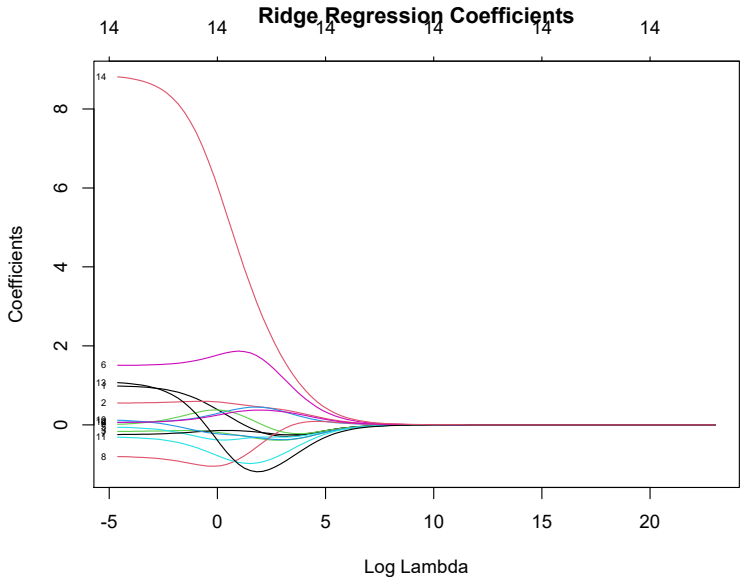

Figura 15: Grafico coefficienti Ridge Regression

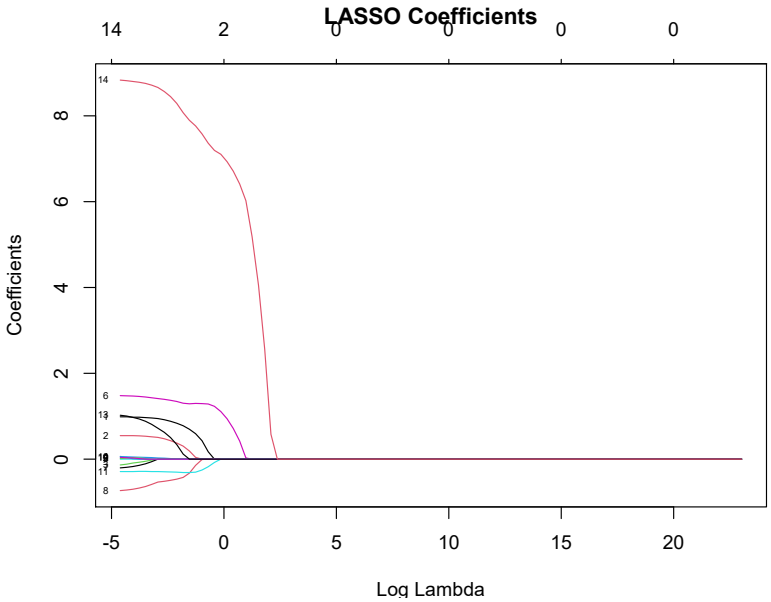

Figura 16: Grafico coefficienti LASSO

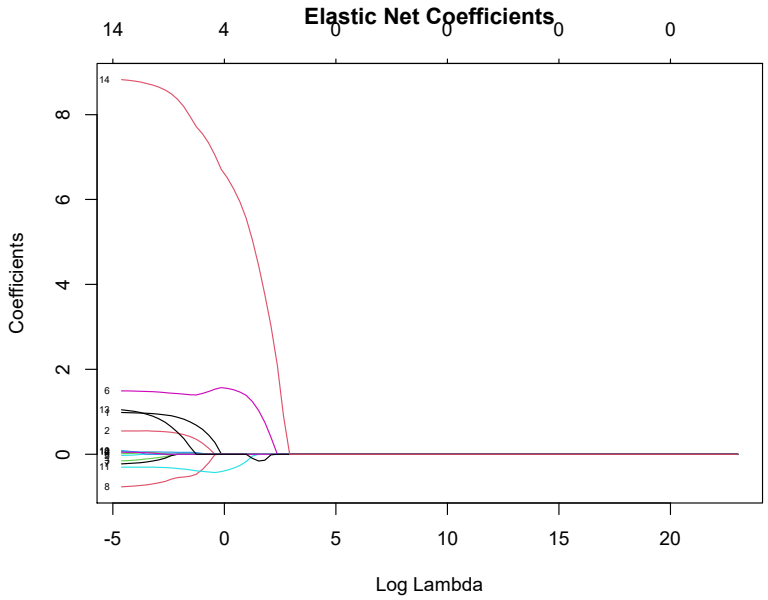

Figura 17: Grafico coefficienti Elastic Net