Project Work Data Mining

Università degli Studi di Milano Bicocca

Corso di Laurea in Statistica e Gestione delle Informazioni

Gruppo di Lavoro

Keita Jacopo Viganò
Sara Borello
Lorenzo Giuliano

Indice

1 Introduzione al progetto **2**
1.1 Scelta delle Variabili 3

2 Missing Data **4**
2.1 Conteggio valori mancanti per ogni variabile 4
2.2 Eliminazione delle osservazioni problematiche 4

3 Optimal Grouping e Trasformazioni delle X **5**

4 Primo FIT del modello **8**
4.1 Output 8
4.2 Problematiche 9

5 Multicollinearità **10**
5.1 VIF e TOL 10

6 Linearità **12**
6.1 Box-Cox 12
6.2 Trasformazioni GAM 13
6.2.1 Spline 13
6.2.2 LOESS 15
6.2.3 Confronto modello con Spline vs Loess 16

7 Outliers, Valori Anomali e Punti Influenti **18**

8 Eteroschedasticità **19**
8.1 Standard errors Robusti di White 19

9 Model Selection **21**

10 Bootstrap **22**
10.1 Modello Iniziale vs Modello Finale 23
10.2 Interpretazione dei coefficienti 24

11 Modello logistico **25**
11.1 Quasi Separation e Nuovo Fit 26

1 Introduzione al progetto

L'industria edilizia ha un impatto significativo sull'ambiente, contribuendo in modo sostanziale alle emissioni globali di CO2. Comprendere i fattori alla base di queste emissioni è fondamentale per formulare strategie efficaci di riduzione. Oltre al consumo energetico quotidiano, anche la progettazione, costruzione e demolizione degli edifici giocano un ruolo nell'emissione di CO2.

OBIETTIVO

L'obiettivo del presente studio è l'elaborazione di un modello lineare robusto, fondamentale per il suo utilizzo prospettico come strumento di previsione su set di dati non ancora esaminati. Il modello è mirato specificatamente a discernere e quantificare l'effetto di molteplici variabili sulle quantità emesse di CO2 da un edificio abitativo.

DESCRIZIONE DEL DATASET

Il set di dati impiegato in questo studio è stato acquisito dal Portale OpenData della Regione Lombardia, in particolare dal Database CENED, il quale si concentra sulla Certificazione Energetica degli Edifici. Tale archivio dati, comprendente gli Attestati di Prestazione Energetica, è composto da 1,52 milioni di osservazioni organizzate in 40 variabili. A causa dell'ampio volume del dataset, si è proceduto con un campionamento sistematico che ha ridotto il numero di osservazioni a 202019 unità. L'indagine è stata successivamente focalizzata esclusivamente sugli edifici a destinazione residenziale continuativa (variabile DESTINAZIONE_DI_USO= E.1(1) per DPR 412/1993), escludendo di conseguenza quelli ad uso pubblico, il che ha comportato un'ulteriore diminuzione delle osservazioni, stabilizzandosi su un totale di 16.858 unità.

VARIABILE TARGET

La variabile target selezionata è rappresentata dalle emissioni di CO2, quantificate annualmente. In particolare, questa variabile, viene misurata in $KgCO_{2eq}/m^2anno$, offrendo un indicatore del rilascio di gas serra per unità di superficie. Questa metrica facilita il confronto tra edifici di diverse dimensioni. Dalla Figura 1 si nota che le emissioni si concentrano tra 0 e 100 $KgCO_{2eq}/m^2anno$

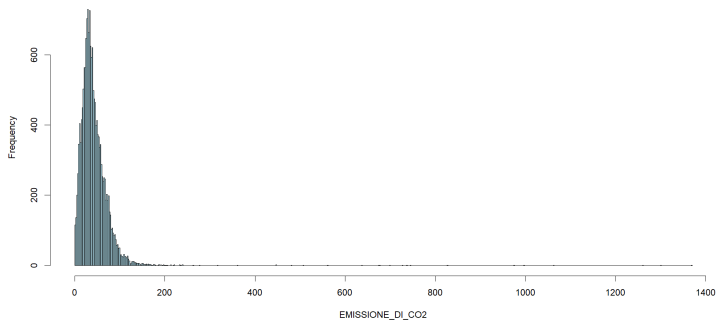

Figura 1: Distribuzione della Variabile Emissione di CO2

1.1 Scelta delle Variabili

Nella fase di sviluppo del modello, non si è fatto ricorso all'intera gamma di 40 variabili disponibili; si è piuttosto limitata la scelta a 14 variabili ritenute più rilevanti, le cui descrizioni sono approfondite nella sezione Allegati Un'analisi grafica di queste variabili in relazione alla variabile dipendente (figure 2 e 3), ha rivelato anomalie che suggeriscono la presenza di errori di digitazione, attribuibili alla scala eccessivamente ampia. Tali osservazioni verranno eliminate nei prossimi step.

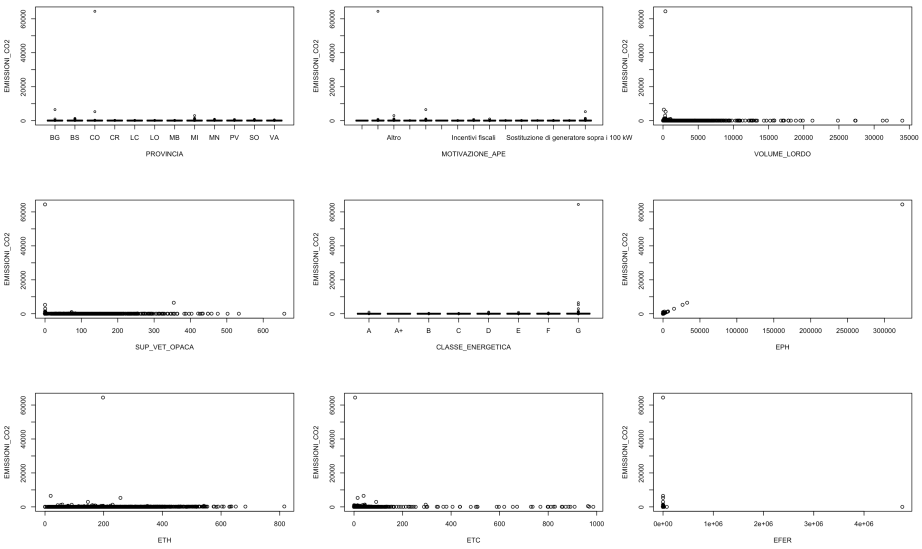

Figura 2: Distribuzione Covariate vs Variabile Emissione di CO2

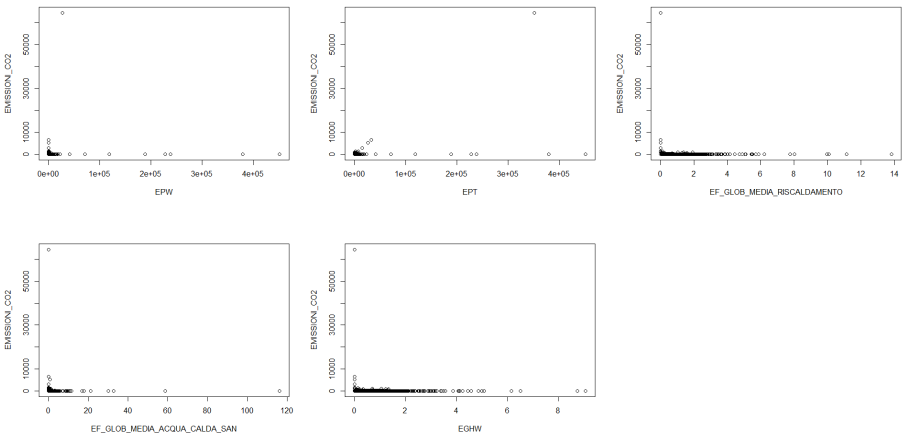

Figura 3: Distribuzione Covariate vs Variabile Emissione di CO2

2 Missing Data

Prima di passare all'analisi vera e propria, è necessaria una fase in cui bisogna pulire il dataset selezionando soltanto le osservazioni che potranno effettivamente essere usate nell'analisi.

2.1 Conteggio valori mancanti per ogni variabile

Dall'esame della Figura 4, si constata che le variabili MOTIVAZIONE_APE e SUPERFICIE_VETRATA_OPACA presentano pochi valori mancanti, rispettivamente 2 e 4, suggerendo una possibile assenza casuale di dati; pertanto, tali osservazioni verranno escluse.

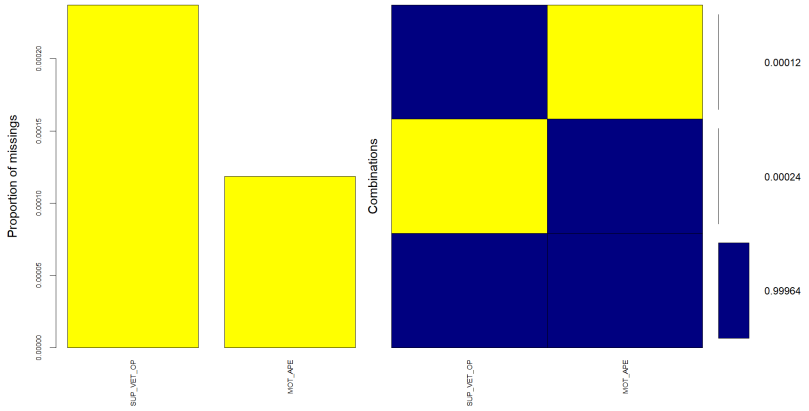

Figura 4: Conteggio dei dati mancanti per variabile

2.2 Eliminazione delle osservazioni problematiche

Dalle analisi grafiche iniziali è emerso che il dataset contiene alcuni valori anomali. Questi sono presumibilmente dovuti a errori durante il processo di inserimento dei dati da parte degli utenti. Le Tabelle 1 e 2 elencano le istanze in cui si manifestano tali irregolarità; specificatamente, per la variabile EFER, che misura il contributo energetico degli impianti a fonti energetiche rinnovabili, sono stati rilevati valori negativi,i quali non sono plausibili. Analogamente per la variabile CO2, che presenta ordini di grandezza errati.

Tabella 1: Pratiche APE anomale

NUM_OSS	EFER
3749	-103.2057
6453	-27.8379
12586	-33.7199
9227	4761936,9772

Tabella 2: Pratiche APE anomale

NUM_OSS	CO2
3229	5285.031
7935	2965.068
10104	6497.558
16509	64406.222

3 Optimal Grouping e Trasformazioni delle X

Una volta trattati i valori mancanti, prima di procedere con l'analisi, è cruciale apportare modifiche ad alcune variabili di tipo categorico, poiché presentano un numero eccessivo di livelli. Le variabili in questione sono CLASSE_ENERGETICA, PROVINCIA e MOTIVAZIONE_APE come riportato nella Figura 5

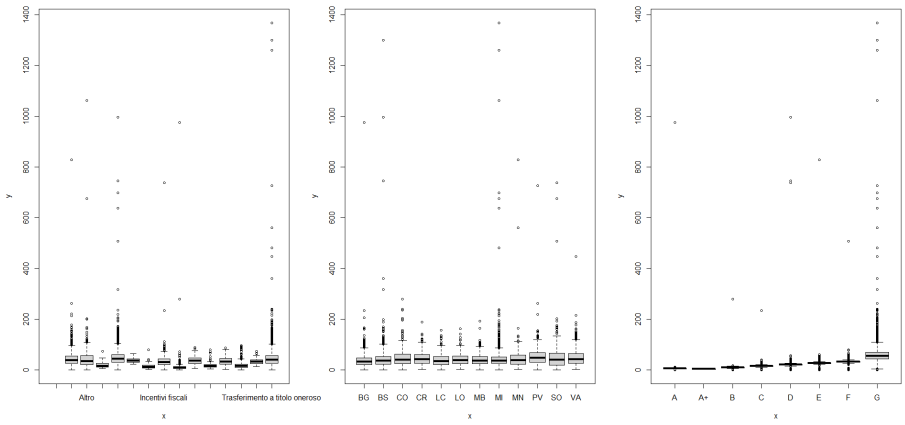

Figura 5: Covariate prima del raggruppamento

La variabile **PROVINCIA**, grafico centrale nella Figura 5, è composta da 12 livelli. Pertanto, si è scelto di procedere con una categorizzazione in 3 gruppi distinti, basata sulla densità abitativa in quanto ques'ultima può riflettere meglio l'intensità dell'uso energetico e le relative necessità infrastrutturali. La nuova partizione è mostrata nella Figura 6

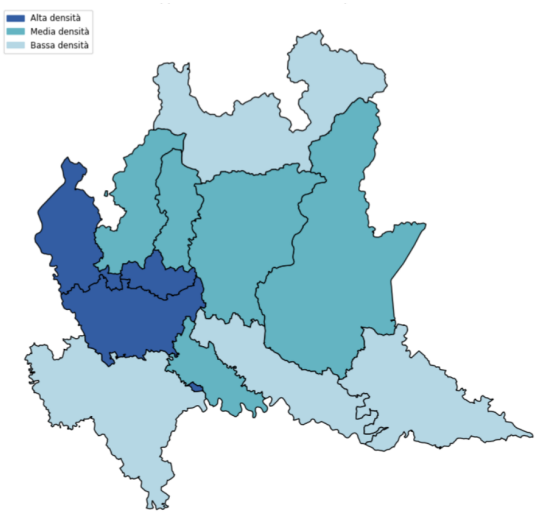

Figura 6: Province per densità abitativa

La variabile **CLASSE_ENERGETICA**, inizialmente articolata in 8 livelli come mostrato nel grafico a destra nell Figura 5 è stata riorganizzata in 3 gruppi distinti come mostrato nella Tabella 3

Tabella 3: Classe Energetica

Gruppo	Descrizione	Conteggio
1	Classe Alta	332
2	Classe Media	3678
3	Classe Bassa	12834

La variabile **MOTIVAZIONE_APE** ha 14 livelli, per cui è stata adottata la metodologia dell'optimal grouping per aggregare categorie simili, secondo la similarità in media rispetto alla variabiel target riducendo i livelli superflui. A seguito di tale applicazione, sono stati delineati 6 cluster, visibili nel dendogramma illustrato nella Figura 7 e specinficati dettagliatamente nella Tabella 4 .

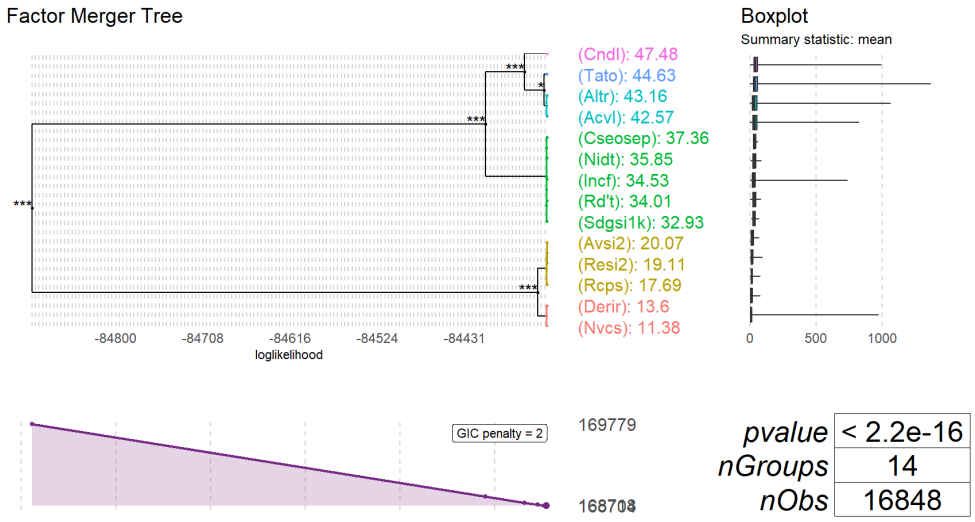

Figura 7: Optimal Grouping

Tabella 4: Motivazioni per l'apertura della pratica APE

Gruppo	Descrizione	Conteggio
1	Nuove costruzioni	1017
2	Interventi strutturali maggiori	487
3	Riqualificazione energetica	950
4	Altro	2219
5	Trasferimento a titolo oneroso	6976
6	Contratto di locazione	5195

Il grafico 8 mostra che l'optimal grouping è stato effettuato efficacemente e che vi è una discreta differenza tra i gruppi rispetto alla variabile dipendente.

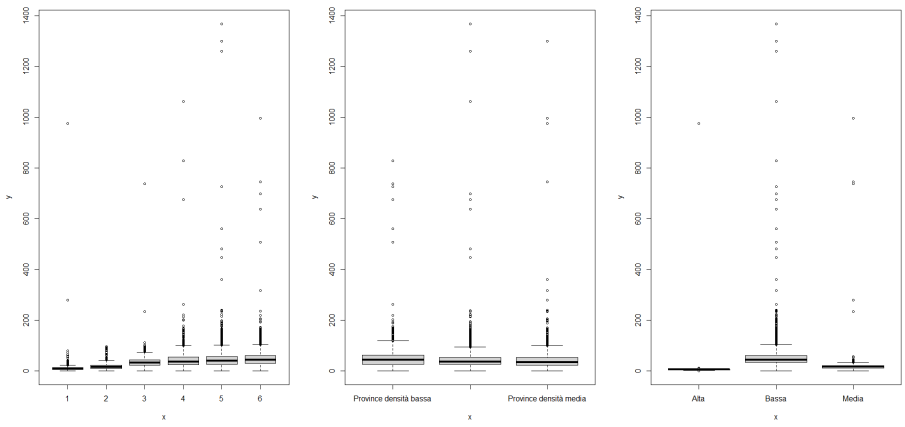

Figura 8: Covariate dopo il raggruppamento

Dopo aver trattato i valori mancanti, corretto gli errori di battitura e codificato le variabili categoriche che presentano un numero di livelli gestibile, si conclude la fase di pre-elaborazione del dataset. Tale conclusione consente di procedere con la prima stima del modello. In particolare, la figura 9 espone in maniera chiara la distribuzione delle covariate continue in relazione al target, offrendo una rappresentazione più intellegibile rispetto a quella proposta nella sezione antecedente. Si osserva la presenza di alcune relazioni non lineari tra le covariate e la variabile dipendente, le quali potrebbero generare problematiche relative all'assunzione di linearità all'interno del modello.

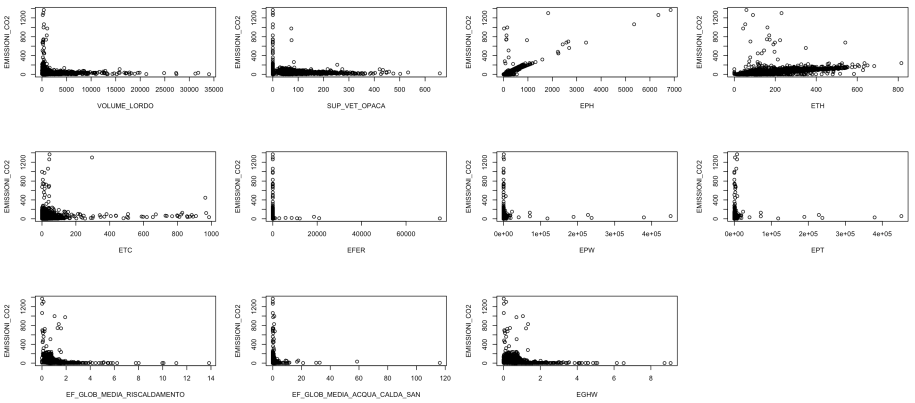

Figura 9: Covariate Continue dopo il Pre-Processing

4 Primo FIT del modello

Il presente paragrafo descrive l'implementazione iniziale di un modello di regressione lineare. L'intento è quello di accertare il soddisfacimento delle ipotesi necessarie per assicurare la robustezza del modello.

4.1 Output

La Figura 10 mostra i risultati del primo fit del modello, si nota che la variabile EPT è stata esclusa a causa di singolarità, suggerendo una possibile correlazione perfetta con altre variabili nel modello, inoltre si nota la presenza di standard error bassi. Dall'analisi, le variabili che risaltano per la loro significatività sono: PROVINCIA_DENSITA_ALTA, CLASSE_ENERGETICA_BASSA, CLASSE_ENERGETICA_MEDIA, EPH, ETH, ETC, e EF_GLOB_MEDIA_RISCALDAMENTO. Il modello è stato correttamente configurato su 16.844 osservazioni.

Variable		N	Estimate	p
PROVINCIA	Province densità bassa	2632	Reference	
	Province densità elevata	8148	-0.80 (-1.61, 0.02)	0.05
	Province densità media	6064	-0.41 (-1.25, 0.43)	0.34
MOTIVAZIONE_APE_REDUCED	1	1017	Reference	
	2	487	-0.65 (-2.67, 1.38)	0.53
	3	950	0.54 (-1.24, 2.33)	0.55
	4	2219	0.17 (-1.39, 1.73)	0.83
	5	6976	-0.03 (-1.48, 1.42)	0.97
	6	5195	0.11 (-1.38, 1.60)	0.88
VOLUME_LORDO		16844	0.00 (-0.00, 0.00)	0.58
SUPERFICIE_VETRATA_OPACA		16844	0.00 (-0.00, 0.01)	0.48
CLASSE_ENERGETICA	Alta	332	Reference	
	Bassa	12834	-3.46 (-6.09, -0.83)	0.01
	Media	3678	-2.14 (-4.59, 0.31)	0.09
EPH		16844	0.22 (0.21, 0.22)	<0.001
ETH		16844	-0.01 (-0.02, -0.01)	<0.001
ETC		16844	0.01 (0.00, 0.02)	0.03
EFER		16844	-0.00 (-0.00, 0.00)	0.93
EPW		16844	-0.00 (-0.00, 0.00)	0.82
EPT		16844		
EGHW		16844	-1.10 (-2.54, 0.33)	0.13
EF_GLOB_MEDIA_RISCALDAMENTO		16844	1.20 (-0.14, 2.54)	0.08
EF_GLOB_MEDIA_ACQUA_CALDA_SAN		16844	-0.02 (-0.28, 0.23)	0.86

Figura 10: Fit Modello Completo

Tabella 5: Summary Fit Modello Completo

Misure di Adattamento	**Valori**
Residual standard error	18.24 on 16824 df
Multiple R-squared	0.7616
Adjusted R-squared	0.7613
F-statistic:	2824 on 19 and 16824 DF
p-value:	$<$ 2.2e-16

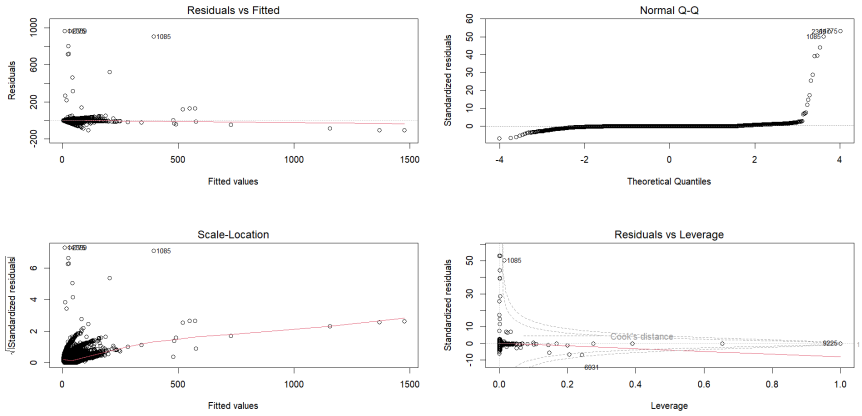

Figura 11: Grafico dei residui

4.2 Problematiche

Dalle diagnostiche dei residui presentate nella Figura 11, emergono le seguenti problematiche:

- **MULTICOLINEARITA'**: La variabile EPT non è stata inclusa nel fit del modello a causa di un problema di singolarità, la quale si verifica quando c'è una perfetta collinearità, o una correlazione estremamente alta, tra due o più variabili predittive.

- **LINEARITA'**:L'esame del grafico che confronta i residui con i valori adattati (Fitted Values) rivela una certa non linearità. Questa circostanza implica la possibile necessità di trasformare la variabile dipendente e le covariate. Se trascurata, tale problematica potrebbe compromettere la correttezza e l'efficienza delle stime dei coefficienti, rendendo l'interpretazione dell'inferenza problematica.

- **ETEROSCHEDASTICITA'**: Nell'osservare il grafico dei residui standardizzati in rapporto ai valori interpolati, si nota che la LOESS non ha pendenza nulla. Questo suggerisce l'assenza di omoschedasticità e la presenza di correlazione tra i residui e i valori interpolati. Di fronte a tale scenario, si riscontrano problematiche relative a standard error non adeguati e a un'errata inferenza sui parametri del modello.

- **PUNTI INFLUENTI**: Nel grafico dei Residui Studentizzati vs Leverage, si osservano alcune osservazioni con una distanza di Cook notevolmente alta. Contemporaneamente, è evidente la presenza di outlier sulle y a causa di alcuni residui studentizzati elevati; mentre sul lato delle x, sono presenti leverage notevoli, segnalando l'esistenza di punti di leva.

5 Multicollinearità

La multicollinearità può distorcere le stime dei parametri, rendendole inaffidabili, in particolare la multicollinearità perfetta rende la stima impraticabile poiché la matrice X'X non è invertibile, come visto con la variabile EPT. Nonostante il primo modello non abbia mostrato standard error elevati, ulteriori controlli sono necessari, utilizzando VIF, TOL e χ^2.

5.1 VIF e TOL

Nell'analisi di multicollinearità (Tabella 6), la variabile EPT mostra un VIF infinito, rivelando chiaramente il suo legame con EPW ed EPH. Questa collinearità è evidenziata anche dal valore zero del TOL per le medesime variabili. Tuttavia, esaminando globalmente le altre covariate, i valori VIF si mantengono prossimi all'unità e non oltrepassano il valore critico di 10. Analogamente, i valori di TOL rimangono superiori a 0.3, indicando l'assenza di problemi di multicollinearità.

Tabella 6: Valori di VIF e TOL per le variabili analizzate.

Variabile	VIF	TOL
VOLUME_LORDO	1.0080	0.9920
SUPERFICIE_VETRATA_OPACA	1.0056	0.9944
EPH	Inf	0.0000
ETH	2.1411	0.4671
ETC	1.0086	0.9914
EFER	1.0023	0.9977
EPW	Inf	0.0000
EPT	Inf	0.0000
EGHW	1.9794	0.5052
EF_GLOB_MEDIA_RISCALDAMENTO	1.8528	0.5397
EF_GLOB_MEDIA_ACQUA_CALDA_SAN	1.1091	0.9016

Per verificare la presenza di una connessione tra variabili categoriche, si utilizza il χ^2 normalizzato. Osservando la Tabella 7, si evidenzia che il valore di quest'ultimo è molto basso per tutte le combinazioni possibili tra le covariate categoriche; pertanto, si deduce che non vi è connessione.

Tabella 7: Test del χ^2

Row	Column	χ^2	df	p.value	n	u1	u2	χ^2_N
PROV	APE	138.571	10	0	16848	2	5	0.0041
PROV	CE	114.733	4	0	16848	2	2	0.0034
APE	CE	6044.914	10	0	16848	5	2	0.1795

Proseguendo con l'esclusione della variabile EPT, si è giunti alla tabulazione dei risultati riportati nella Tabella 8, dove si evidenzia una situazione lontana dalla presenza di multicollinearità.

Tabella 8: Valori di VIF e TOL dopo l'eliminazione di EPT

Variabile	VIF	TOL
VOLUME_LORDO	1.0080	0.9921
SUPERFICIE_VETRATA_OPACA	1.0059	0.9942
EPH	2.080	0.4529
ETH	2.1415	0.4670
ETC	1.0086	0.9914
EFER	1.0014	0.9986
EPW	1.0079	0.9922
EGHW	1.9762	0.5060
EF_GLOB_MEDIA_RISCALDAMENTO	1.8485	0.5410
EF_GLOB_MEDIA_ACQUA_CALDA_SAN	1.1091	0.9016

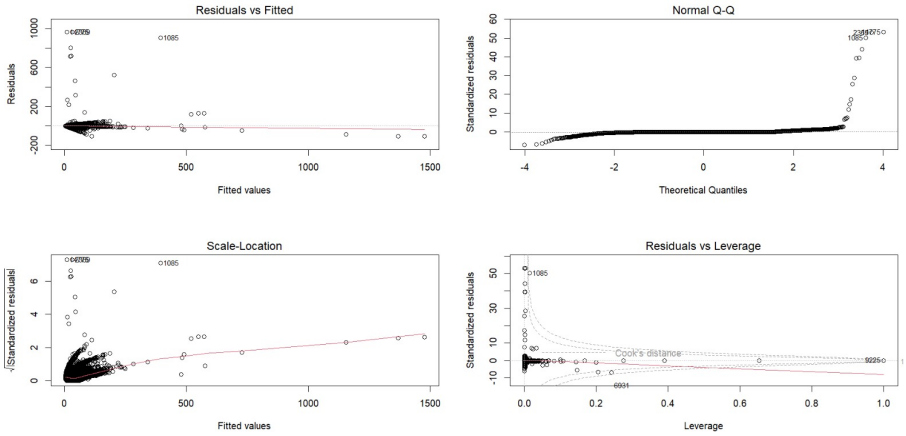

Figura 12: Diagnostiche Fit 2

Osservando la figura 12, non si nota una differenza sostanziale rispetto alle diagnostiche presentate a seguito del primo fit del modello (figura 11).

6 Linearità

6.1 Box-Cox

Nell'effettuare la trasformazione Box-Cox, al fine di minimizzare l'RMSE, è stato inizialmente aggiunto un valore unitario alla variabile target per evitare complicazioni derivanti dall'applicazione del logaritmo. Dalla Figura 13, si evince che il valore di λ che minimizza RMSE è 0.34. Di conseguenza, verrà effettuata una trasformazione logaritmica sulla variabile delle emissioni di CO2.

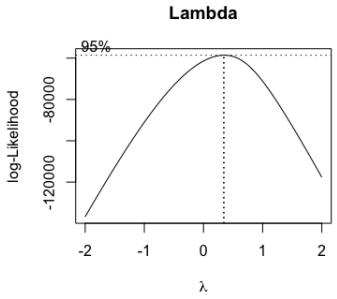

Figura 13: Lambda BoxCox

A seguito dell'applicazione della trasformazione suggerita da Box-Cox (Tabella 9) si procede con un nuovo fit del modello ottenendo le diagnostiche nella Figura 14, la quale suggerisce la necessità di ulteriori trasformazioni sulle variabili indipendenti per correggere la lieve non linearità osservata.

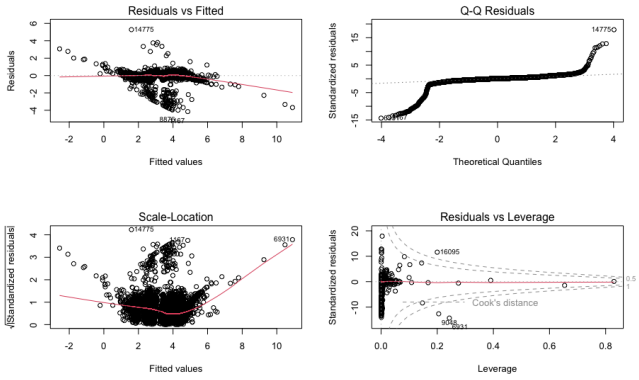

Figura 14: Residui dopo BoxCox

Tabella 9: Summary di y e log(y_pos)

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
y	0.00	23.66	36.06	41.95	54.20	1368.00
log(y_pos)	0.000	3.205	3.613	3.544	4.011	7.222

6.2 Trasformazioni GAM

6.2.1 Spline

L'uso delle spline ha prodotto risultati significativi, come dimostrato nelle tabelle ANOVA per gli effetti parametrici e non parametrici. In particolare, esaminando la Tabella 11, che presenta gli effetti non parametrici, si nota che le variabili influenzate dagli effetti delle spline, ovvero s(EPH), s(ETH), e s(EF_RISC), insieme a s(EGHW) e s(EF_ACQ), mostrano una significatività statistica molto alta, con valori p quasi nulli. Questo suggerisce che la relazione tra queste variabili e la variabile dipendente è complessa e non lineare. Pertanto, si rende necessaria l'adozione di trasformazioni adeguate per modellare con precisione tali non linearità nell'analisi.

Tabella 10: ANOVA for Parametric Effects

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PROV	2	44.71	22.35	365.0199	**<2e-16 *****
APE	5	1057.33	211.47	3452.9096	**<2e-16 *****
s(VOLUME)	1	0.60	0.60	9.7161	0.00183 **
s(SUP_VET)	1	5.55	5.55	90.5476	< 2e-16 ***
CE	2	1240.72	620.36	10129.4931	**<2e-16 *****
s(EPH)	1	2176.86	2176.86	35544.6265	< 2e-16 ***
s(ETH)	1	187.73	187.73	3065.3698	< 2e-16 ***
s(ETC)	1	0.03	0.03	0.5075	0.47624
s(EFER)	1	0.07	0.07	1.1665	0.28015
s(EPW)	1	0.01	0.01	0.1161	0.73333
s(EGHW)	1	80.72	80.72	1318.0438	< 2e-16 ***
s(EF_RISC)	1	132.25	132.25	2159.4039	< 2e-16 ***
s(EF_ACQ)	1	0.07	0.07	1.0970	0.29494
Residuals	16788	1028.14	0.06		

Tabella 11: ANOVA for Nonparam Effects

Effect	Npar Df	Npar F	Pr(F)
(Intercept)			
PROV			
APE			
s(VOLUME)	3.0	3.77	0.0101624 *
s(SUP_VET)	9.0	0.71	0.7015605
CE			
s(EPH)	3.0	541.84	< 2.2e-16 ***
s(ETH)	3.0	789.07	< 2.2e-16 ***
s(ETC)	3.0	2.54	0.0546395 .
s(EFER)	3.0	0.51	0.6760455
s(EPW)	3.1	1.57	0.1923828
s(EGHW)	3.0	43.95	< 2.2e-16 ***
s(EF_RIS)	3.0	457.90	< 2.2e-16 ***
s(EF_ACQ)	3.0	6.48	0.0002229 ***

I grafici mostrati nella Figura 15 illustrano le stime delle funzioni spline per ogni predittore rispetto al predatore corrispondente. Per ogni predittore è descritta una specifica trasformazione, indicata nella parte superiore.

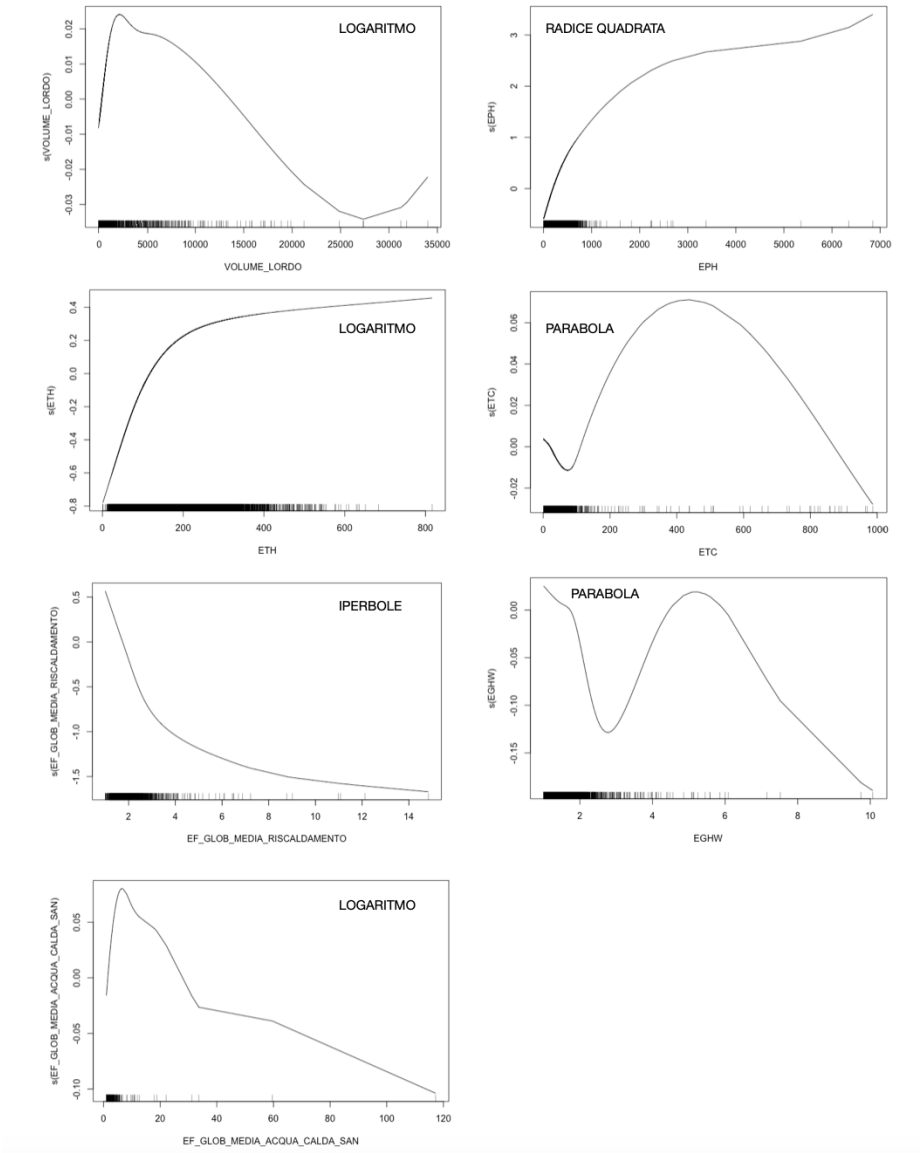

Figura 15: Trasformazioni Spline

6.2.2 LOESS

L'uso del metodo LOESS con uno span di 0.2 e 0.5 fornisce un mezzo per identificare trasformazioni ottimali per i dati. Dall'applicazione di LOESS con questi parametri di span, emergono delle trasformazioni suggerite che sono visualizzate nella Figura 16

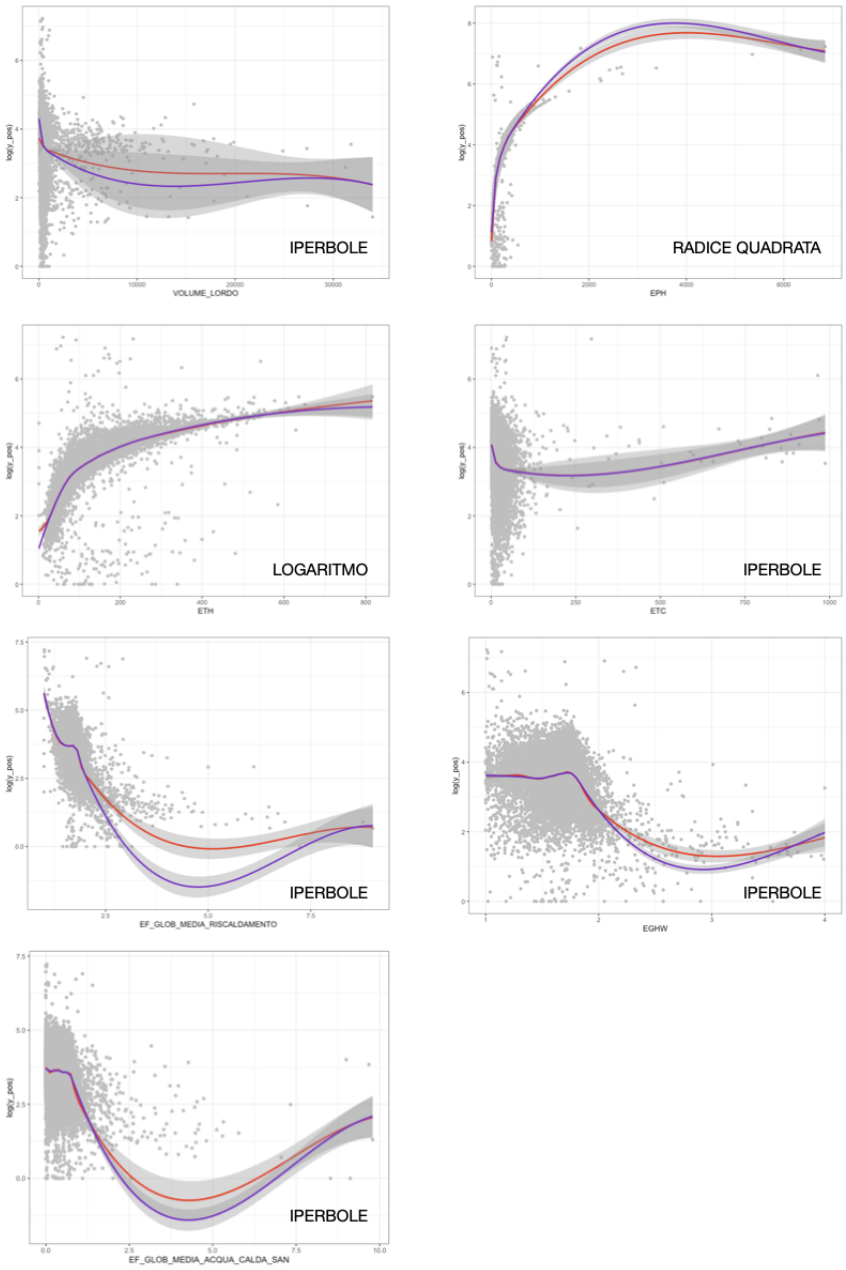

Figura 16: Trasformazioni LOESS

6.2.3 Confronto modello con Spline vs Loess

Esaminando la Tabella 12, si evidenzia come il modello che incorpora le trasformazioni suggerite dalle Spline, superi in performance di linearità il modello LOESS. Di conseguenza, per l'elaborazione del modello aggiornato si prevede l'impiego di trasformazioni suggerite dalle spline, come descritto nel paragrafo 6.2.1.

Tabella 12: Modello LOESS vs Modello Spline

Model	Res.Df	RSS	Reset	P-VALUE Reset Test
Modello Spline	16824	1038.0	57.24	4.058×10^{-14}
Modello Loess	16822	1037.2	65.67	5.695×10^{-16}

La Tabella 13 fornisce un sommario della significatività di varie variabili trasformate, indicando che l'uso delle trasformazioni suggerite dalle spline contribuisce a una migliore rappresentazione delle relazioni non lineari tra le variabili.

Tabella 13: Fit Trasformazioni Spline

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	-2.3045	0.0714	-32.27	< 2e-16
PA	1.550e-02	5.676e-03	2.73	6.319e-03
PM	9.595e-04	5.839e-03	0.164	8.6947e-01
APE2	-1.485e-02	1.414e-02	-1.050	2.9364e-01
APE3	-7.364e-03	1.269e-02	-0.580	5.6168e-01
APE4	-1.276e-02	1.127e-02	-1.133	2.5729e-01
APE5	-4.771e-03	1.058e-02	-0.451	6.5213e-01
APE6	-8.502e-04	1.094e-02	-0.078	9.3803e-01
log(VOLUME)	1.102e-02	3.165e-03	3.481	5.010e-04
SUP_VET	8.014e-05	5.260e-05	1.524	1.2759e-01
CB	3.925e-02	2.200e-02	1.784	7.4435e-02 .
CM	-1.311e-02	1.878e-02	-0.698	4.8506e-01
$I(EPH^{(0.5)})$	4.239e-02	1.549e-03	27.367	< 2e-16
log(ETH)	6.580e-01	1.251e-02	52.597	< 2e-16
$I(ETC^2)$	3.700e-08	1.756e-07	0.211	8.3311e-01
ETC	7.586e-06	1.286e-04	0.059	9.5296e-01
EFER	1.487e-06	3.015e-06	0.493	6.2190e-01
EPW	-1.422e-07	3.459e-07	-0.411	6.8089e-01
$I(EGHW^2)$	-2.536e-03	2.913e-03	-0.871	3.8384e-01
EGHW	-4.157e-02	1.971e-02	-2.109	3.4930e-02
I((1/EF_RISC))	3.458	7.543e-02	45.844	< 2e-16
log(EF_ACQ)	8.454e-02	1.510e-02	5.598	2.2e-08

Analizzando le diagnostiche dei residui a seguito del fit del nuovo modello (Figura 17) emerge un miglioramento nella linearità, ma permangono segni di eteroschedasticità, indicando che il modello può essere ulteriormente ottimizzato.

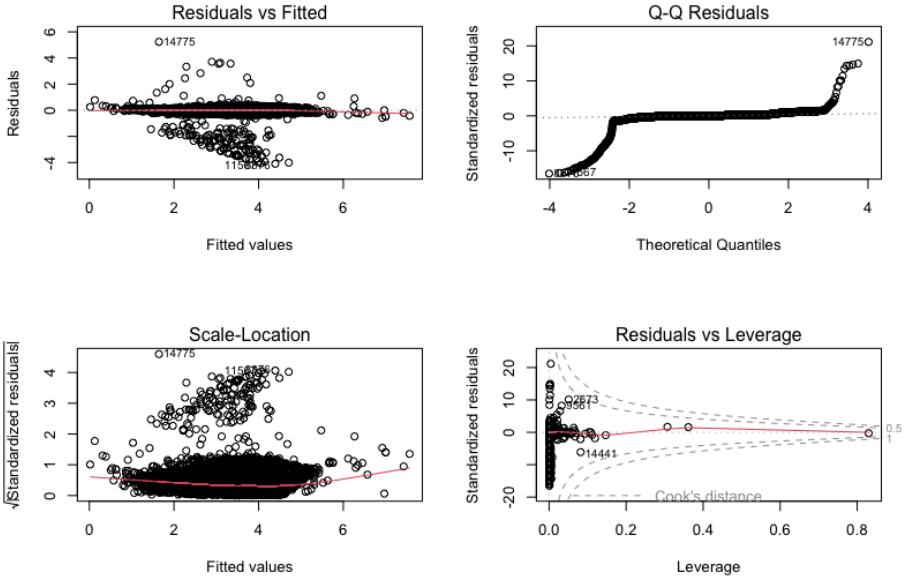

Figura 17: Diagnostiche con trasformazioni Spline

Per controllare se l'ipotesi di linearità è stata soddisfatta, si effettua il RESET test. La tabella 14 indica che l'assunzione di linearità non è stata ancora raggiunta. Una possibile spiegazione di questo risultato potrebbe essere l'influenza di valori anomali sul modello e la presenza di eteroschedasticità.

Tabella 14: Risultati del RESET test

RESET	df1	df2	p-value
57.239	1	16821	4.058×10^{-14}

7 Outliers, Valori Anomali e Punti Influenti

In questa sezione, si attua l'eliminazione dei punti influenti che esercitano una distorsione significativa sul modello. Osservando la Figura 18, si rileva la presenza di alcuni punti che potrebbero alterare in modo rilevante le stime del modello. Per mitigare questo problema, si interviene escludendo i valori distorsivi attraverso l'utilizzo dei DFITTS, ottenendo così un nuovo fit aggiornato.

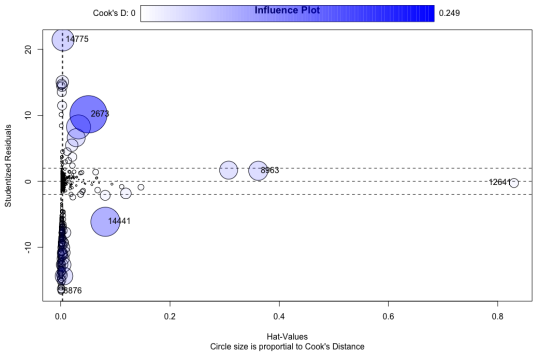

Figura 18: Valori influenti

I risultati delle diagnostiche sono rappresentati nella Figura 19, la quale mostra un avanzamento nella linearità, evidenziato da una diminuzione della densità della nuvola di punti sottostante la retta. Applicando il RESET test (p-value= 0.6521), si procede con l'accettazione dell'ipotesi nulla. Questo significa che non sono presenti deviazioni significative dalla linearità ipotizzata nelle assunzioni del modello, confermando così il beneficio derivante dall'eliminazione di determinate osservazioni.

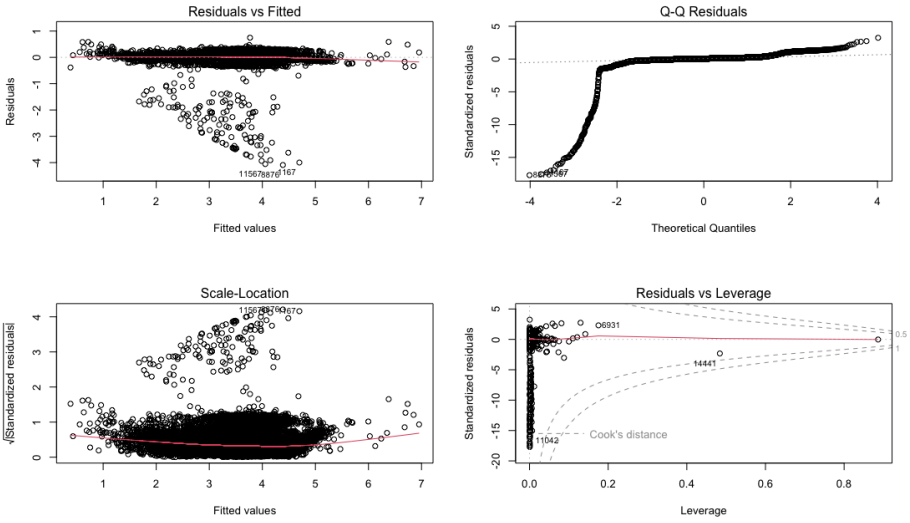

Figura 19: Diagnostiche Fit 5

8 Eteroschedasticità

La presenza di eteroschedasticità dei residui è stata verificata attraverso l'applicazione del Breusch-Pagan Test e del White Test, i quali hanno portato al rigetto di H0, come evidenziato in Tabella 15.

Tabella 15: Brush Pagan Test

Fit	Value	Degrees of Freedom	p-value
fit5	243.39	21	$< 2.2 \times 10^{-16}$
fit5	152.94	1	$< 2.2 \times 10^{-16}$

8.1 Standard errors Robusti di White

In seguito a questo risultato, si è proceduto con la stima dei parametri impiegando gli standard error robusti di White (Tabella 16).

Tabella 16: Robust Standard Errors

Term	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-2.7406e+00	1.0912e-01	-25.1154	< 2.2e-16
PA	2.3270e-02	6.2588e-03	3.7180	0.0002015
PM	5.7393e-03	7.0556e-03	0.8134	0.4159792
APE2	-1.6638e-02	1.3981e-02	-1.1901	0.2340410
APE3	-2.9362e-02	1.3604e-02	-2.1584	0.0309099
APE4	-3.0142e-02	1.1416e-02	-2.6403	0.0082915
APE5	-2.3352e-02	9.7288e-03	-2.4003	0.0163923
APE6	-2.1636e-02	1.0013e-02	-2.1609	0.0307147
log(VOLUME)	9.7989e-03	2.7970e-03	3.5033	0.0004607
SUP_VET	3.9477e-05	2.8000e-05	1.4099	**0.1585975**
CB	-4.9654e-02	2.7641e-02	-1.7964	0.0724512
CM	-8.0672e-02	2.3411e-02	-3.4458	0.0005707
I($EPH^{0.5}$)	3.1530e-02	2.6441e-03	11.9246	< 2.2e-16
log(ETH)	7.5201e-01	2.1005e-02	35.8022	< 2.2e-16
I(ETC^2)	4.3667e-08	1.3020e-07	0.3354	**0.737332**
ETC	-1.9625e-05	1.0678e-04	-0.1838	**0.854174**
EFER	1.5995e-06	9.4490e-07	1.6927	0.090525
EPW	-4.3960e-07	3.8622e-07	-1.1382	**0.255053**
I($EGHW^2$)	-3.2290e-02	1.4495e-02	-2.2277	0.0259121
EGHW	-2.3301e-02	4.6079e-02	-0.5057	0.6130995
I((1/EF_RISC))	3.8390	1.2590e-01	30.4935	< 2.2e-16
log(EF_ACQ)	1.1937e-01	3.1176e-02	3.8288	0.0001293

Dall'analisi è emerso che le variabili SUP_VET, EPW ed ETC non risultano essere statisticamente significative quando si adotta un approccio di inferenza robusta. Al fine di tentare una risoluzione del problema di eteroschedasticità, è stato adottato un modello rivisto che esclude le variabili sopracitate.

Tuttavia, come si può osservare dalle diagnostiche del nuovo fit del modello (Figura 20) rispetto al fit precedente (Figura 19) e dalla successiva applicazione del Breusch-Pagan Test (Tabella 17), non si registra un miglioramento sostanziale della condizione di eteroschedasticità

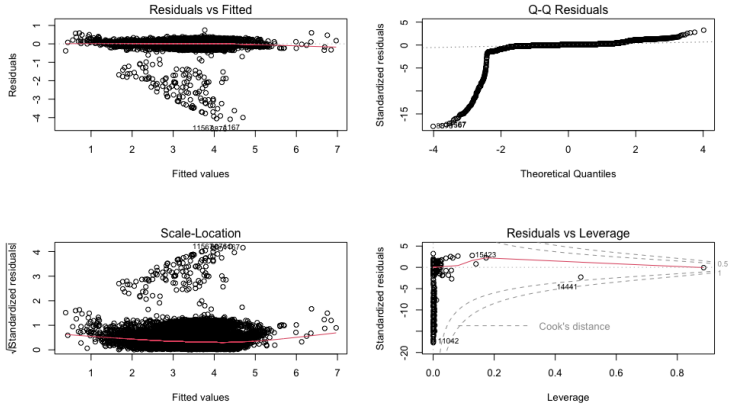

Figura 20: Diagnostiche Fit 6

Tabella 17: Brush Pagan Test Fit 6

Fit	BP value	Degrees of Freedom	p-value
fit6	244.04	17	$< 2.2 \times 10^{-16}$

Per avere una conferma definitiva, i due modelli (il modello completo e il modello ridotto) vengono confrontati utilizzando ANOVA, come mostrato nella Tabella 18. Da questo confronto emerge che l'ipotesi nulla viene accettata, il che indica che entrambi i modelli forniscono una spiegazione equivalente dei dati. Di conseguenza, si preferisce adottare il modello ridotto per il principio della parsimonia.

Tabella 18: ANOVA Modello Ridotto vs Modello Completo

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	16768	895.93				
2	16764	895.87	4	0.062894	0.2942	0.8819

9 Model Selection

Per scegliere il modello più adatto, si adotta prima di tutto il criterio AIC, che individua come miglior sottoinsieme quello indicato nella tabella 19 , con un valore AIC pari a -49155.96. Successivamente, si procede con la stima del modello includendo le variabili facenti parte di questo miglior sottoinsieme, osservando che la variabile EFER è stata omessa.

Tabella 19: Model selection con AIC

	Df	Sum of Sq	RSS	AIC
- EGHW	1	0.058	896.01	-49157
- APE	5	0.502	896.45	-49157
<none>			895.95	-49156
+ EFER	1	0.016	895.93	-49154
- log(VOLUME)	1	0.590	896.54	-49147
- PROVINCIA	2	1.566	897.52	-49131
- I(EGHWˆ2)	1	1.932	897.88	-49122
- CE	2	2.914	898.86	-49105
- log(EF_ACQU)	1	3.551	899.50	-49092
- I((EPH)ˆ(0.5))	1	21.702	917.65	-48756
- I((1/EF_RISC))	1	147.461	1043.41	-46600
- log(ETH)	1	185.257	1081.21	-46003

Infine, si procede utilizzando il metodo drop-1 (Tabella 20) applicato al Fit del nuovo modello, da cui si evince che tutte le variabili risultano essere significative. Si decide di includere anche la variabile APE poiché il suo p-value è inferiore a 0.10.

Tabella 20: Drop 1 Fit 7

Term	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			895.95	-49156		
PROV	2	1.566	897.52	-49131	14.6517	4.390e-07
APE	5	0.502	896.45	-49157	1.8786	0.0944354 .
log(VOLUME)	1	0.590	896.54	-49147	11.0378	0.0008946
CE	2	2.914	898.86	-49105	27.2685	1.502e-12
$I(EPH^{0.5})$	1	21.702	917.65	-48756	406.1788	<2.2e-16
log(ETH)	1	185.257	1081.21	-46003	3467.3594	<2.2e-16
$I(EGHW^2)$	1	1.932	897.88	-49122	36.1601	1.855e-09
EGHW	1	0.058	896.01	-49157	1.0808	0.2985336
I((1/EF_RISC))	1	147.461	1043.41	-46600	2759.9487	<2.2e-16
log(EF_ACQ)	1	3.551	899.50	-49092	66.4624	3.816e-16

10 Bootstrap

Dall'utilizzo del metodo bootstrap si ottengono gli intervalli di confidenza degli stimatori β. Analizzando le Figure 21 e 22, si valuta se ciascun β_{mle} rientra all'interno dell'intervallo di confidenza bootstrap. In particolare, si osserva che solo PROVINCIA a densità media, APE2, EGHW e CLASSE ENERGETICA Bassa includono lo zero nei loro intervalli di confidenza, indicando che solo alcuni livelli delle variabili categoriche presentano queste caratteristiche. Tuttavia, nel caso di EGHW, sebbene il suo effetto lineare non sia significativo, la trasformazione parabolica mostra significatività. Questo implica che potrebbe essere opportuno conservare queste variabili nel modello, in quanto contribuiscono a spiegare la varianza della variabile risposta.

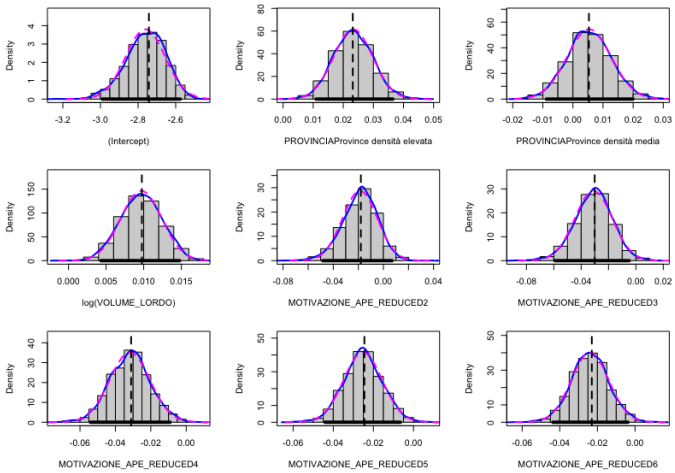

Figura 21: Intervalli di confidenza Bootstrap

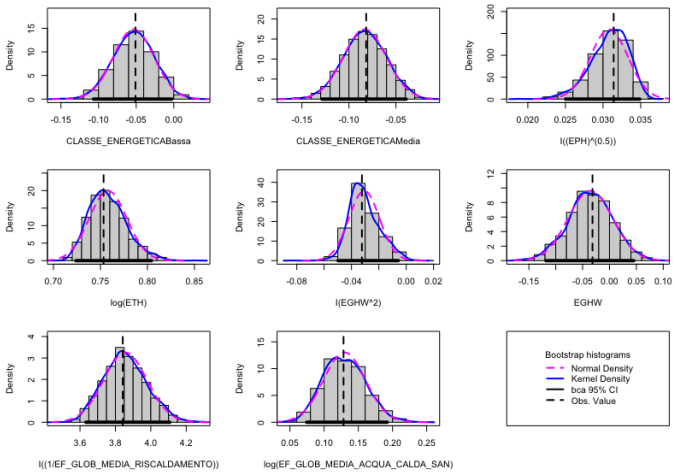

Figura 22: Intervalli di confidenza Bootstrap

10.1 Modello Iniziale vs Modello Finale

Confrontando il modello iniziale con l'ultima versione testata, si osserva un evidente miglioramento: dalla Figura 23, si può notare come la linearità sia notevolmente perfezionata, confermata anche dal reset test. Dal confronto tra le due analisi diagnostiche (Figure 24, 25) emerge anche un miglioramento dell'eteroschedasticità, sebbene non sia stata raggiunta una soluzione definitiva.

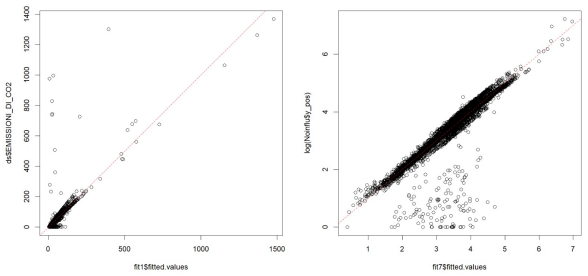

Figura 23: Confronto Linearità

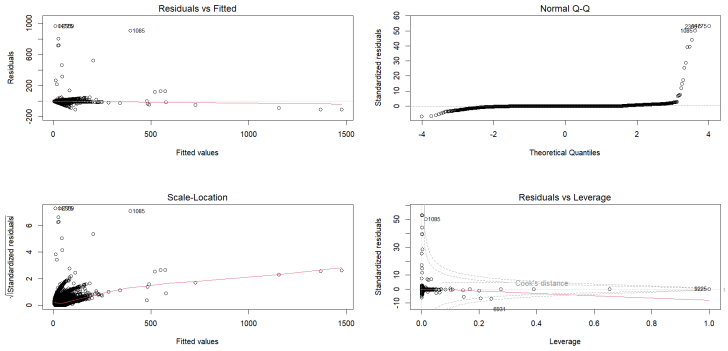

Figura 24: Diagnostiche Fit 1

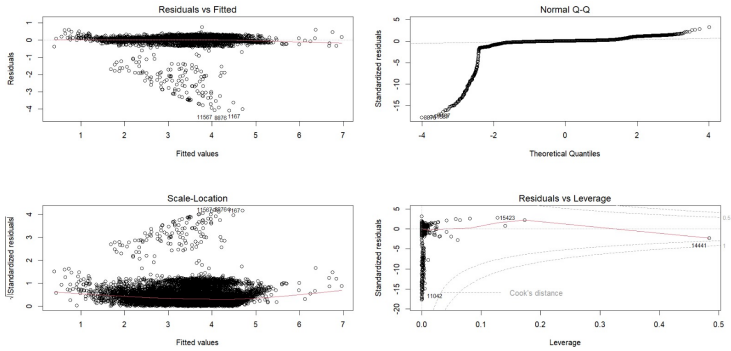

Figura 25: Diagnostiche Fit 7

Infine, analizzando il test complessivo sulle ipotesi alla base del modello, si conferma che l'ipotesi di linearità è stata confermata e l'eteroschedasticità è stata in gran parte mitigata.

Tabella 21: Test globale sulle assunzioni del modello

Test	Value	p-value	Decision
Global Stat	1.461e+07	0.0000	Assumptions NOT satisfied!
Skewness	3.341e+05	0.0000	Assumptions NOT satisfied!
Kurtosis	1.428e+07	0.0000	Assumptions NOT satisfied!
Link Function	1.861e-01	0.6662	Assumptions acceptable.
Heteroscedasticity	1.014e+02	0.0000	Assumptions NOT satisfied!

10.2 Interpretazione dei coefficienti

L'analisi ha confermato che le emissioni di CO2 degli edifici sono influenzate da una serie di fattori, tra cui il tipo di provincia, la finalità dell'attestato di prestazione energetica (APE), la classe energetica dell'edificio, l'efficienza dei sistemi di riscaldamento e produzione di acqua calda, nonché il fabbisogno energetico annuo e il volume dell'edificio.

Variable		N	Estimate	p
PROVINCIA	Province densità bassa	2614		Reference	
	Province densità elevata	8130		0.02 (0.01, 0.03)	<0.001
	Province densità media	6042		0.01 (-0.00, 0.02)	0.287
MOTIVAZIONE_APE_REDUCED	1	1001		Reference	
	2	485		-0.02 (-0.04, 0.01)	0.204
	3	947		-0.03 (-0.05, -0.01)	0.013
	4	2214		-0.03 (-0.05, -0.01)	0.004
	5	6957		-0.02 (-0.04, -0.00)	0.018
	6	5182		-0.02 (-0.04, -0.00)	0.035
log(VOLUME_LORDO)		16786		0.01 (0.00, 0.02)	<0.001
CLASSE_ENERGETICA	Alta	315		Reference	
	Bassa	12810		-0.05 (-0.09, -0.01)	0.020
	Media	3661		-0.08 (-0.12, -0.05)	<0.001
I((EPH)^(0.5))		16786		0.03 (0.03, 0.03)	<0.001
log(ETH)		16786		0.75 (0.73, 0.78)	<0.001
I(EGHW^2)		16786		-0.03 (-0.04, -0.02)	<0.001
EGHW		16786		-0.02 (-0.06, 0.02)	0.299
I((1/EF_GLOB_MEDIA_RISCALDAMENTO))		16786		3.84 (3.70, 3.98)	<0.001
log(EF_GLOB_MEDIA_ACQUA_CALDA_SAN)		16786		0.12 (0.09, 0.15)	<0.001

0 1 2 3

Figura 26: Diagnostiche Fit 7

Tabella 22: Interpretazione coefficienti variabili categoriche ($exp($beta)-1)*100)

Variabile	Coefficiente
(Intercept)	-93.55824
PROVINCIA Province densità elevata	2.34225
PROVINCIA Province densità media	0.53256
MOTIVAZIONE_APE_REDUCED2	-1.78204
MOTIVAZIONE_APE_REDUCED3	-2.96614
MOTIVAZIONE_APE_REDUCED4	-3.06359
MOTIVAZIONE_APE_REDUCED5	-2.41770
MOTIVAZIONE_APE_REDUCED6	-2.27025
CLASSE_ENERGETICABassa	-4.96970
CLASSE_ENERGETICAMedia	-7.84251

11 Modello logistico

Il settore delle costruzioni residenziali ha registrato un calo nell'intensità delle emissioni di CO2, scendendo da 43 a 40 kg per metro quadrato tra il 2015 e il 2021[1]. Tale misura è stata adottata come parametro di soglia per categorizzare le emissioni di CO2 delle abitazioni, distinguendo tra quelle che sono conformi alle tendenze globali di riduzione delle emissioni (1) e quelle che non lo sono (0). Questo criterio mira a riconoscere le abitazioni che si allineano agli obiettivi globali di decarbonizzazione previsti per il 2050, una priorità sottolineata nel corso dell'incontro di Sharm El Sheikh il 9 novembre 2022.

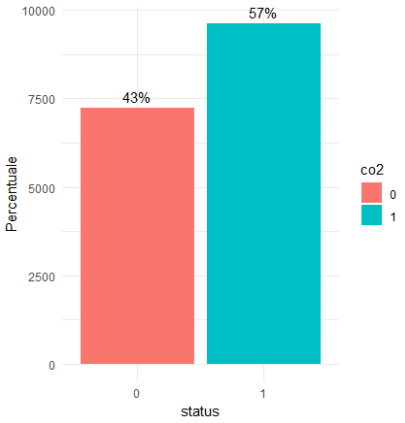

Figura 27: Distribuzione di CO2

Previsti	Osservati	
	0	1
0	7020	132
1	213	9479

Figura 28: Matrice di confusione

Tabella 23: Fit modello logistico

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	24.8578	4.4268	5.62	0.0000
PA	0.0070	0.1304	0.05	0.9570
PM	0.2686	0.1367	1.97	0.0494
APE2	-0.3523	1.0605	-0.33	0.7397
APE3	-0.5000	0.9879	-0.51	0.6128
APE4	-0.5634	0.9792	-0.58	0.5650
APE5	-0.4681	0.9742	-0.48	0.6309
APE6	-0.5994	0.9749	-0.61	0.5387
CB	21.8337	1.5577	14.02	0.0000
CM	16.8002	1.5329	10.96	0.0000
log(VOLUME)	-0.0297	0.0688	-0.43	0.6660
$I(EPH^{0.5})$	-2.8638	0.1350	-21.21	0.0000
log(ETH)	-0.7148	0.8210	-0.87	0.3839
$I(EGHW^2)$	1.1589	0.6445	1.80	0.0722
EGHW	-0.4729	0.8242	-0.57	0.5661
I((1/EF_RISC))	-3.4036	3.9779	-0.86	0.3922
log(EF_ACQ)	-0.5040	0.3629	-1.39	0.1649

[1]CO2 emissions from buildings and construction hit new high, leaving sector off track to decarbonize by 2050: UN

Il modello non presenta problemi quali varianza nulla o multicollinearità; tuttavia, esaminando le stime dei coefficienti per i livelli della variabile CLASSE ENERGETICA, si notano valori molto elevati sintomo di quasi separation (Tabella 23).

11.1 Quasi Separation e Nuovo Fit

Si osserva nella Tabella 24 che la prevalenza delle abitazioni con una classe energetica elevata-media risulta conforme alle tendenze di riduzione di CO2 (indicato con il valore 1), ciò significa che la variabile classe energetica con livelli alto e medio fornisce un'informazione quasi deterministica sulla probabilità di conformità dati. Pertanto si decide di ristimare il modello escludendo tale variabile (Figura 29) .

Tabella 24: Quasi separation classe energetica

	0	1
Alta	1	331
Bassa	7221	5613
Media	11	3667

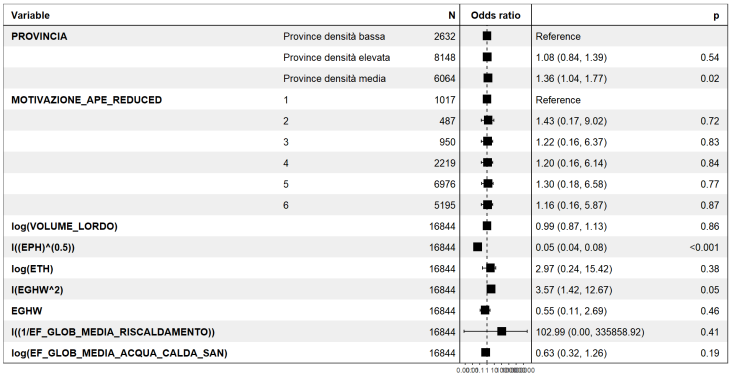

Figura 29: Fit modello logistico senza classe energetica

Con un livello di accuratezza pari a 0,9793, il modello dimostra un adattamento efficace ai dati, come illustrato nella Tabella 25. Questa elevata accuratezza, che rimane invariata anche dopo la rimozione della variabile affetta da quasi-separation, può sembrare in prima istanza un indicatore positivo della bontà del modello. Tuttavia, un valore così elevato richiede un'indagine ulteriore, in quanto potrebbe nascondere problemi come l'overfitting

Tabella 25: Matrice di confusione

Previsti	Osservati	
	0	1
0	7020	135
1	213	9476