UNIVERSITÀ DI PISA

Corso di laurea magistrale in Artificial Intelligence and data engineering

Statistica

Progetto 1

Giuseppe Martino

22 novembre 2020

Indice

1 Introduzione

L'obiettivo di questa analisi è quello di fornire un modello a una agenzia immobiliare internazionale o a un privato locatore che permetta di fornire un supporto per la decisione del prezzo dell'affitto mensile di un appartamento di fascia media , in base allo standard di vita e ai prezzi medi di beni e servizi della città in cui si trova l'immobile.

Il dataset analizzato contiene il prezzo di alcuni prodotti, servizi e il salario medio in varie città del mondo. Si è svolta un'analisi regressiva, allo scopo di trovare un modello che riuscisse a ottenere una buona previsione del prezzo di un appartamento, considerato come fattore di uscita del modello e cercando inoltre di capire quanto bene la varianza di questo fattore venisse spiegata dagli altri fattori di ingresso.

2 Dati

2.1 Creazione del dataset

Il dataset è stato creato utilizzando i dati dei report annuali di Deutsche Bank, "Mapping the world's prices" del 2019 e del 2017, disponibili ai seguenti link:

- 2019: https://www.dbresearch.com/PROD/RPS_EN-PROD/PROD0000000000494405/Mapping_the_world%27s_prices_2019.pdf?undefined&realload=hfviQ5u~YRPWnU/whEnJMpzJ8Qk/MNwfHCAuaAWosMv60ZgqmBQXTa2YKHdeY0XkcGJhkG8ikkGd719uofU6Sw==
- 2017: https://www.finews.ch/images/download/Mapping.the.worlds.prices.2017.pdf

In particolare ho utilizzato le colonne '2019' delle figure 2,3,5,6,7,15,17,20,21 del 2019 e le colonne '2017' delle figure 2,3,5,6,7,16,18,21,22 del 2017.
Tra tutte le città indicate nelle varie tabelle del report del 2019, ho escluso le città di San Francisco, Riyadh,Shangai e Dhaka in quanto presentavano dei dati mancanti.

Il dataset dunque, risulta così strutturato dopo la creazione:

Numero di osservazioni:98
Numero dei fattori iniziali: 11

Viene riportata di seguito una descrizione degli attributi della tabella.

- City: Città.
- Year: Anno.
- Salary: Stipendio netto medio mensile.
- Rent: Prezzo dell'affitto mensile di un appartamento di fascia media con 2 camere da letto.
- Weekend Getaway: Prezzo per una vacanza di un weekend, calcolata nel seguente modo: due notti in un hotel 5 stelle, due pasti in un pub per due, due cene al ristorante per due, due pinte di birra, 4 litri di acqua,acquisto di un paio di jeans e un paio di scarpe.

- Cheap date: Prezzo per un appuntamento economico, calcolato nel seguente modo: corsa in taxi, pranzo/cena per due in un pub, drink, due biglietti al cinema e un paio di birre.
- Bad Habits: 5 birre e 2 pacchetti di sigarette.
- Ticket Public Transport: Costo di un abbonamento mensile ai mezzi pubblici.
- Gas: prezzo gas (1 litro).
- Gym: prezzo di un abbonamento mensile in una palestra nel distretto Business della città.
- Haircut: prezzo di un taglio da uomo standard

Tutti i prezzi sono espressi in dollari statunitensi (USD).

2.2 Importazione e pulizia del dataset

Dopo aver importato il dataset, ho eliminato i fattori "Year" e "City", considerandoli non rilevanti per l'analisi, dato che gli anni sono solo 2. Il numero di fattori quindi si riduce a 9.

2.3 Correlazione

Per una prima visualizzazione dei dati vediamo il grafico che mostra le correlazioni tra gli attributi.

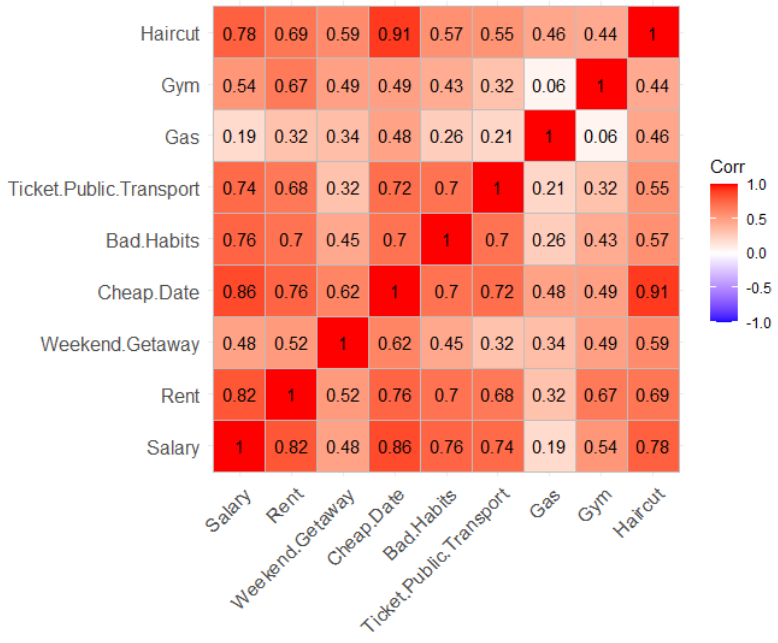

Figura 1: Correlazione tra gli attributi

Analizzando le correlazioni, il fattore scelto come fattore di uscita,"Rent", sembra abbastanza correlato con tutti gli altri fattori, soprattutto con "Salary". Come prevedibile infatti, il canone di affitto di un immobile è più alto nelle città in cui il reddito medio dei lavoratori è più alto. Il fattore meno correlato invece è "Gas".

3 Modello Lineare

3.1 Calcolo del modello

Successivamente è stato calcolato il modello di regressione lineare, considerando il fattore "Rent" come fattore di uscita e tutti gli altri come fattori di ingresso.

```
Call:
lm(formula = Rent ~ ., data = Prices)

Residuals:
    Min      1Q  Median      3Q     Max
-750.94 -186.00  -38.56  132.21 1129.96

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -529.33683  161.06329  -3.287 0.001453 **
Salary                     0.27537    0.07211   3.819 0.000248 ***
Weekend.Getaway            0.08520    0.11703   0.728 0.468510
Cheap.Date                -7.54552    3.36065  -2.245 0.027229 *
Bad.Habits                 4.24932    3.31118   1.283 0.202710
Ticket.Public.Transport    5.21193    1.65436   3.150 0.002221 **
Gas                      390.19145  120.05340   3.250 0.001629 **
Gym                       11.23021    1.82681   6.147 2.18e-08 ***
Haircut                   13.04586    8.18204   1.594 0.114381
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 341.4 on 89 degrees of freedom
Multiple R-squared:  0.807,     Adjusted R-squared:  0.7896
F-statistic:  46.5 on 8 and 89 DF,  p-value: < 2.2e-16
```

Figura 2: Riepilogo modello regressione lineare

Per valutare la bontà del modello, per prima cosa ho considerato il valore della varianza spiegata, il quale è dell'80,7% e della varianza spiegata corretta (78,96%); valori che indicano che i fattori di ingresso spiegano abbastanza bene la variabilità del fattore di uscita.

Per valutare invece il ruolo dei diversi fattori all'interno del modello si pone l'attenzione sui loro p-value. Dall'analisi dei p-value si nota che alcuni fattori hanno un valore abbastanza alto. Per esempio "Weekend.Getaway" ha il valore più alto e, considerando anche la non troppo alta correlazione con il fattore di uscita si può supporre che sia un attributo poco rilevante per il modello. A dispetto della bassa correlazione con il fattore di uscita, il fattore "Gas", avendo un basso p-value, riesce a dare un contributo di spiegazione al modello non irrilevante.

3.2 Riduzione del modello

Sulla base di quanto affermato precedentemente, ho provato a ridurre il modello, eliminando di volta in volta i fattori con p-value alto, con lo scopo di risolvere eventuali allineamenti tra gli attributi e quindi migliorare la robustezza del modello, oltre che per arrivare a un modello che ci permetta comunque di prevedere in maniera affidabile il fattore di uscita, sulla base di un numero minore di fattori in ingresso.

Alla fine di questo procedimento, tenendo conto del decremento della varianza spiegata al decrescere del numero di fattori nel modello, ho ritenuto un buon compromesso il modello contenente solo i fattori "Gas","Gym", "Salary" e "Ticket.Public.Transport" che ci permette di avere un modello con solo quattro fattori di ingresso a fronte di una perdita dell' 1,39% di varianza spiegata e dello 0.54% della varianza spiegata corretta.

A ogni passo, il valore del p-value dell'attributo "Salary" decresce considerevolmente, soprattutto dopo l'eliminazione del fattore "Bad.Habits", per cui è possibile pensare che la capacità di spiegazione di quest'ultimo è anche portata da "Salary",e quindi a un allineamento tra i due fattori, considerando anche la loro buona correlazione.

```
Call:
lm(formula = Rent ~ . - Weekend.Getaway - Bad.Habits - Haircut
    Cheap.Date, data = Prices)

Residuals:
   Min     1Q Median     3Q    Max
-700.0 -218.2  -15.4  118.8 1280.6

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)           -531.96865  144.91296  -3.671 0.000403
Salary                   0.25839    0.04694   5.505 3.25e-07
Ticket.Public.Transport  4.15685    1.44627   2.874 0.005020
Gas                    317.24815   93.40192   3.397 0.001005
Gym                     11.08428    1.73309   6.396 6.35e-09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 345.7 on 93 degrees of freedom
Multiple R-squared:  0.7931,    Adjusted R-squared:  0.7842
F-statistic: 89.13 on 4 and 93 DF,  p-value: < 2.2e-16
```

Figura 3: Riepilogo modello lineare ridotto a quattro fattori di ingresso

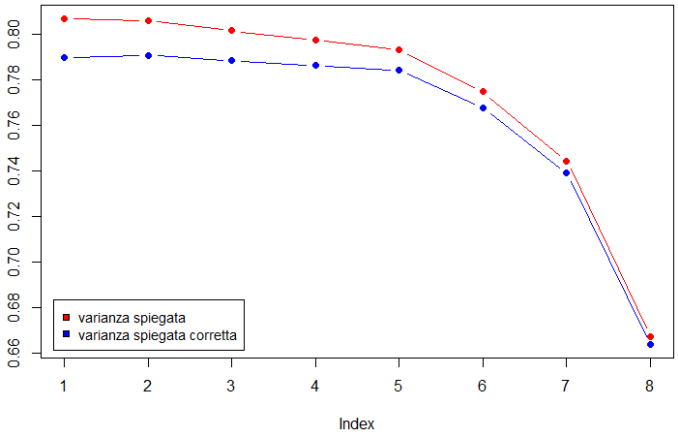

Figura 4: Variazione di varianza spiegata e varianza spiegata corretta al diminuire del numero di fattori di ingresso

4 Modello non lineare

Ho anche valutato l'utilizzo di un modello non lineare per vedere se e come variano le performances. Ho analizzato un modello utilizzando il logaritmo dei dati originali e successivamente ho ridotto il modello con lo stesso procedimento eseguito con il modello lineare. Il risultato è stato un incremento di varianza spiegata che adesso ammonta all'85,95% e la varianza spiegata corretta all'85,35%, riducendo il modello a quattro fattori di ingresso. Confrontando gli attributi rimasti nei due modelli, si nota che nel logaritmico il fattore "Haircut" ha acquisito rilevanza ai danni di "Gas" rispetto al modello lineare.

```
Call:
lm(formula = Rent ~ . - Bad.Habits - Cheap.Date - Weekend.Getaway -
    Gas, data = lPrices)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59526 -0.15338 -0.00028  0.13128  0.68035

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.76447    0.34244   5.153 1.44e-06 ***
Salary                   0.28010    0.06687   4.188 6.38e-05 ***
Ticket.Public.Transport  0.18716    0.06344   2.950 0.004019 **
Gym                      0.42341    0.07079   5.981 4.08e-08 ***
Haircut                  0.26637    0.07105   3.749 0.000308 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2277 on 93 degrees of freedom
Multiple R-squared:  0.8595,    Adjusted R-squared:  0.8535
F-statistic: 142.3 on 4 and 93 DF,  p-value: < 2.2e-16
```

Figura 5: Riepilogo modello logaritmico

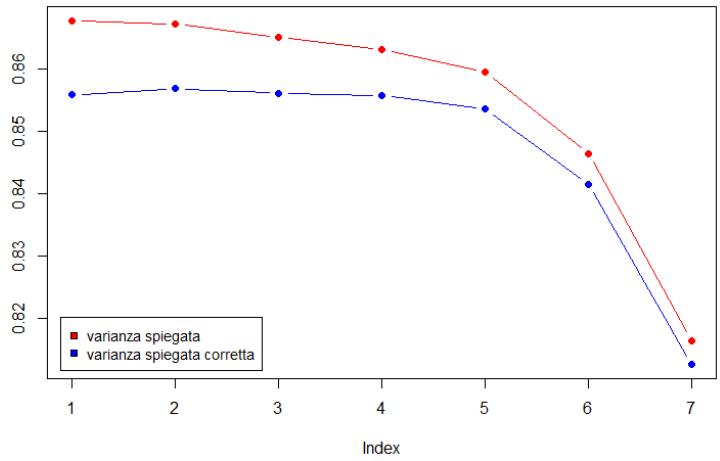

Figura 6: Variazione di varianza spiegata e varianza spiegata corretta al diminuire del numero di fattori di ingresso (Modello logaritmico)

5 Autovalutazione e confronto tra i modelli

Questa fase ha lo scopo di valutare la capacità di predizione dei modelli, confrontandoli tra loro. Inoltre ho cercato di capire se la riduzione dei modelli ha avuto effetti particolari sulle loro capacità di predizione. Dato che le osservazioni non sono molto numerose, non è stato possibile mettere da parte una frazione di dati da usare come validazione dei modelli, perciò il procedimento è stato il seguente: per ogni modello, partendo dal dataset originale sono stati ricavati due set di dati distinti da usare uno come training-set e l'altro come test-set, quest'ultimo composto da 5 osservazioni estratte a caso dal dataset originale e il trainingset, composto dalle restanti osservazioni. Sono state fatte 20 estrazioni per rendere il procedimento più robusto. A ogni iterazione sono state calcolate le previsioni dei modelli calibrati con i training set sui valori dei test set, e infine sono stati calcolati gli errori del modello, calcolando lo scarto tra i valori previsti e quelli effettivi.

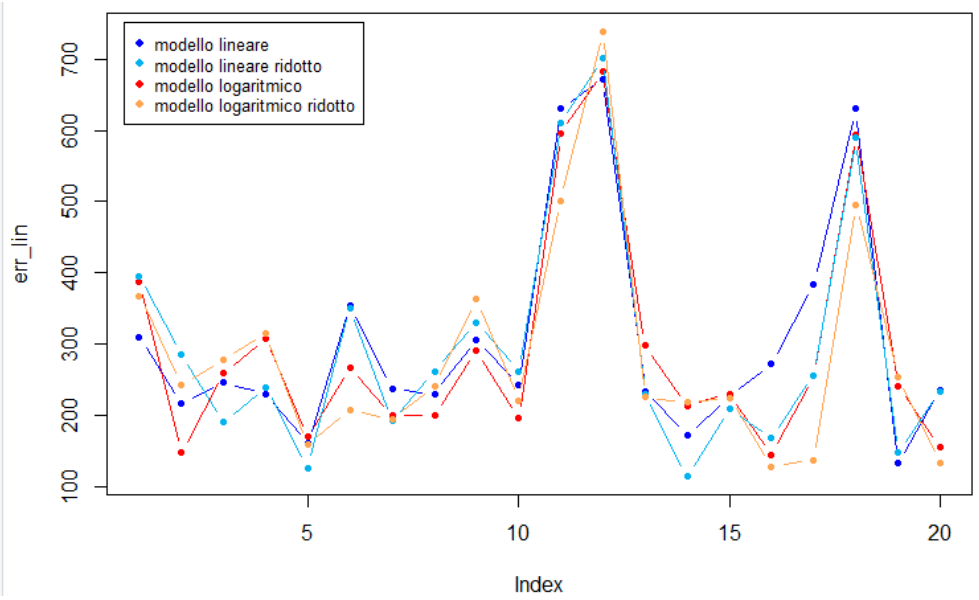

Figura 7: Andamento degli errori

```
>   mean(err_lin)
[1] 305.9446
>   sd(err_lin)
[1] 157.7224
>
>   mean(err_log)
[1] 291.5061
>   sd(err_log)
[1] 156.532
>
>   mean(err_lin_red)
[1] 294.5501
>   sd(err_lin_red)
[1] 163.7779
>
>   mean(err_log_red)
[1] 282.1514
>   sd(err_log_red)
[1] 150.7744
```

Figura 8

La figura 7 mostra l'andamento degli errori medi dei modelli a ogni estrazione e la figura 8 i valori degli errori medi e deviazioni standard dei modelli lineare (err_lin), lineare ridotto (err_lin_red), logaritmico (err_log) e logaritmico ridotto(err_log_red). Si nota come la riduzione sui modelli effettuata ha di poco diminuito l'errore medio in fase di predizione, dovuto probabilmente all'eliminazione di fattori allineati che portano in fase di predizione un errore più rilevante. Confrontando i due modelli, il modello logaritmico ha una capacità di predizione migliore di quello lineare, seppur non di molto. In generale comunque, la differenza tra i modelli è minima e i modelli trovati hanno un errore medio in fase di predizione non così basso, tale da non soddisfare abbastanza lo scopo dell'analisi.

6 Analisi dei residui

É stata svolta un'analisi dei residui per capire quanto i modelli siano riusciti a catturare la struttura del problema.

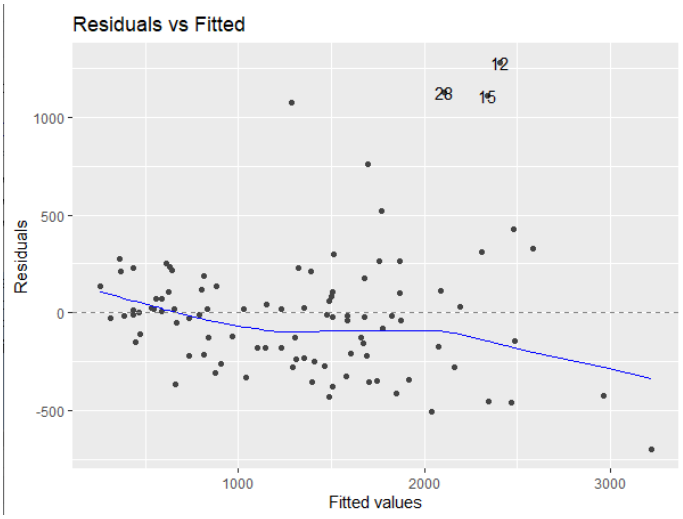

Figura 9: Residui modello lineare

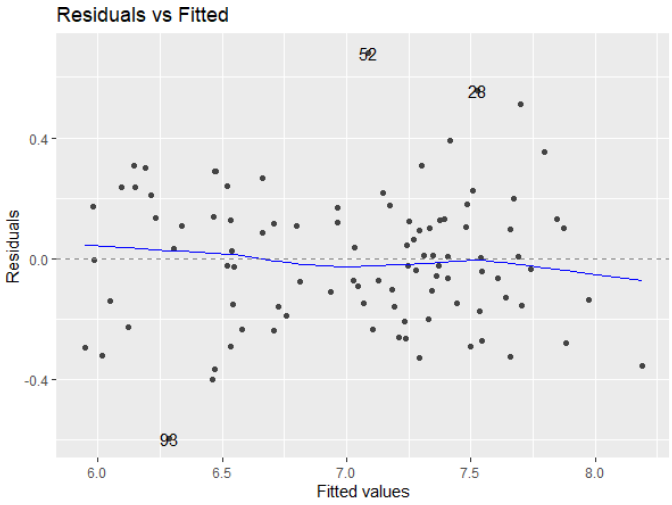

Figura 10: Residui modello logaritmico

Da una prima analisi dei grafici dei residui, notiamo che nel grafico del modello lineare la variabilità dei residui tende ad aumentare all'aumentare del fattore di uscita, e i residui sembrano essere più concentrati in una zona del grafico, a differenza del modello logaritmico, i cui residui sembrano invece disposti nel grafico con più casualità. I residui del modello logaritmico sembrano infatti non presentare struttura e non presentano dipendenza dal fattore di uscita.

A questo punto, per capire la natura del rumore,per ogni modello si è cercato di determinare una distribuzione di probabilità che riuscisse il più possibile a catturare il rumore. Si è tracciato dunque il grafico quantile-quantile per confrontare i quantili empirici dei residui con i quantili teorici della distribuzione normale.

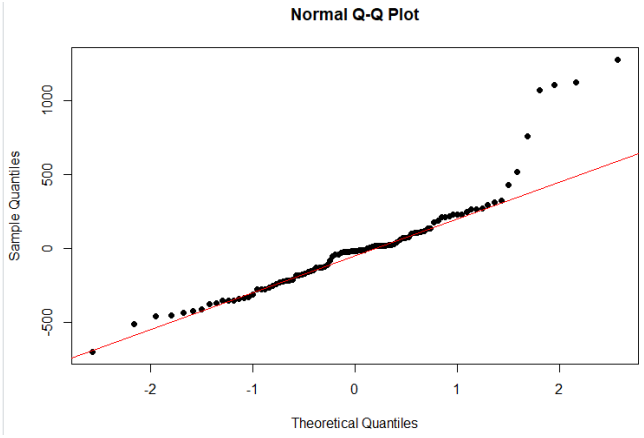

Figura 11: QQ-plot modello lineare

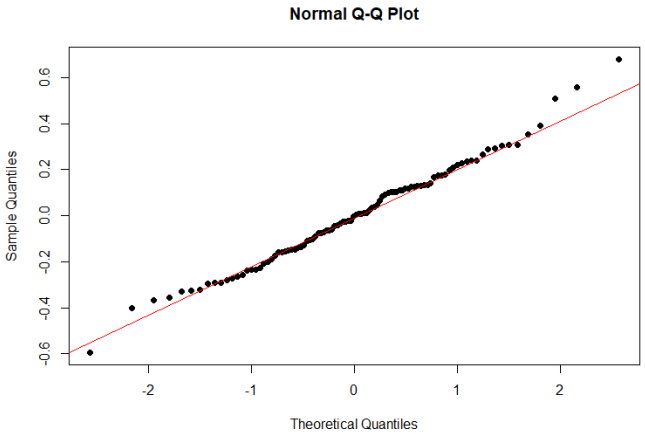

Figura 12: QQ-plot modello logaritmico

Dal QQ-plot del modello lineare, vediamo che i residui si discostano fortemente dalla retta teorica nella coda superiore, a differenza del grafico del modello logaritmico in cui c'è più aderenza alla retta, segno che i residui sono vicini ad avere una struttura gaussiana.

Per controllare e provare a dare ulteriore supporto a questa ipotesi, sono state anche calcolate le funzioni di densità empiriche, confrontando dunque la distribuzione dei residui con una densità di una normale di media e deviazione standard uguali a quelle calcolate per i residui.

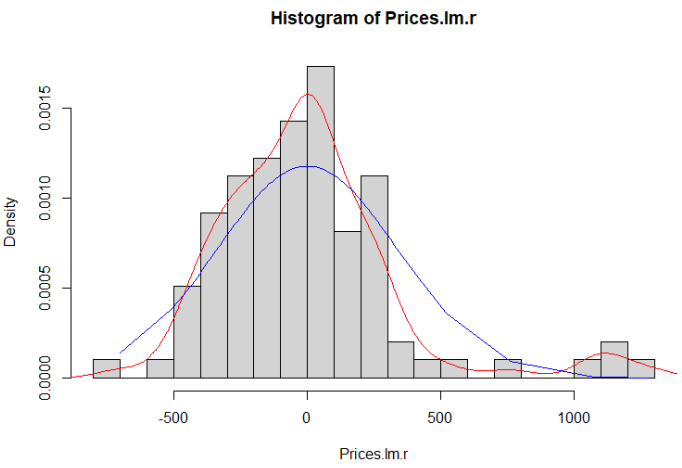

Figura 13: Confronto distribuzioni modello lineare

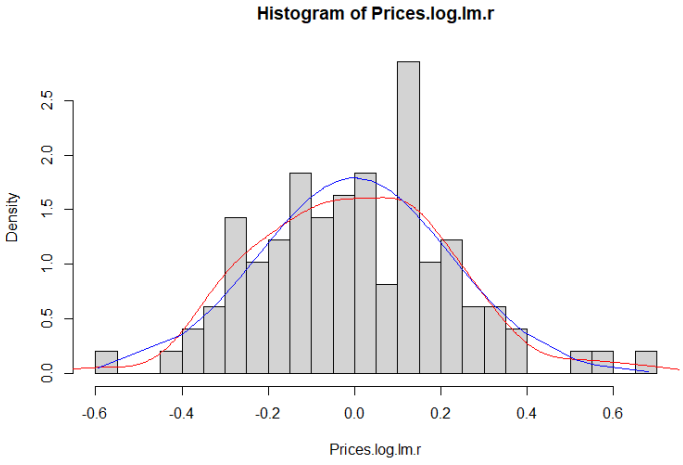

Figura 14: Confronto distribuzioni modello logaritmico

Anche in questo caso la vicinanza tra la curva empirica e quella teorica è migliore nel modello logaritmico.

Infine, una ulteriore prova della struttura gaussiana dei residui ci viene data dallo Shapiro-Wilk test.

Con un p-value di 0.6653 nel modello logaritmico non possiamo rigettare l'ipotesi nulla di normalità dei dati, quindi, considerando tutti gli indizi osservati nell'analisi dei residui, si può affermare che nel modello logaritmico i residui hanno una struttura vicina a una gaussiana, cosa che invece non si può affermare riguardo il modello lineare.

Osservando attentamente i residui, è possibile notare la presenza di valori anomali (outliers). Dopo aver effettuato un controllo sui dati, ho individutato questi outliers nelle città di Hong Kong, Edimburgo e Bangalore. La spiegazione può essere dovuta al fatto che queste città hanno degli affitti parecchio alti (soprattutto Hong Kong) o bassi (Bangalore) e rappresentano quindi un'eccezione nel panorama mondiale.

7 Conclusioni

Posto che il modello logaritmo quindi, riesce meglio a spiegare, seppur di poco, la variabilità del fattore di uscita, come ultimo step ho eliminato gli outliers per provare a migliorare la sua capacità di predizione.

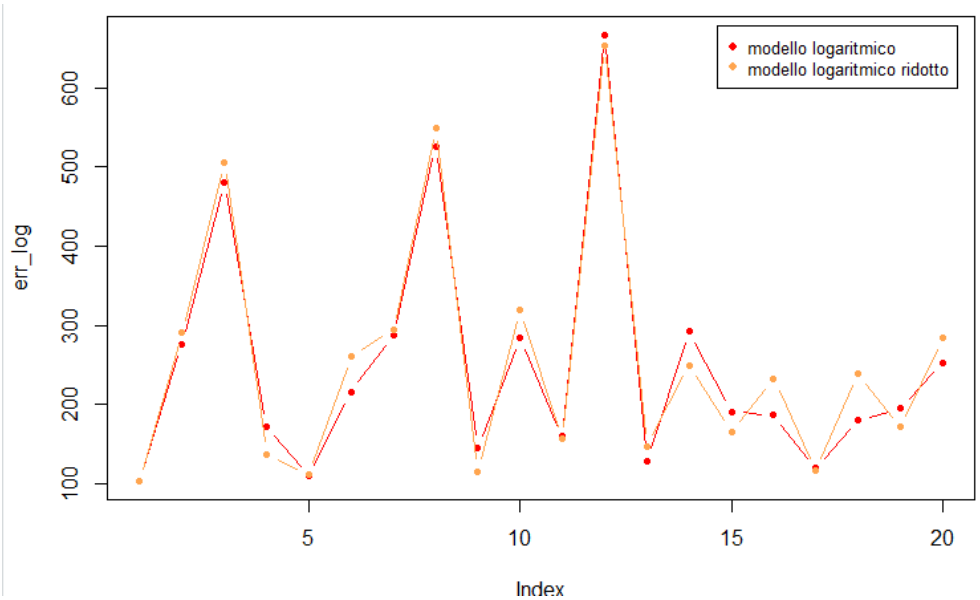

Figura 15: Andamento degli errori

```
>    mean(err_log)
[1] 248.8379
>    mean(err_log_red)
[1] 255.4004
>
>    sd(err_log)
[1] 149.1513
>    sd(err_log_red)
[1] 153.7552
>
>    mean(err_rel)
[1] 0.1893975
>    mean(err_red_rel)
[1] 0.1947361
```

Figura 16

Effettivamente, come si può notare dal valore dell'errore medio rappresentato nelle figure sopra, c'è stato un miglioramento nelle performances, visto un decremento del valore dell'errore medio in fase di predizione.

In conclusione è stato trovato un modello logaritmico che spiega bene la variabilità del fattore di uscita, con dei residui che hanno un aspetto gaussiano. Quindi sono stati ottenuti dei buoni risultati in termini di interpretazione del modello. Per quanto riguarda invece lo scopo predittivo dell'analisi, il modello ha un errore nell'analisi in media di circa il 19%, errore che comunque non è ininfluente nel problema affrontato. Il modello può quindi essere utilizzato dall'agenzia in prima battua per eseguire una prima valutazione generale del prezzo dell'immobile, prezzo che dovrà però poi essere aggiustato considerando anche le caratteristiche della casa, oltre a quella della sua posizione, considerate in questa analisi.