UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Corso di laurea in

DATA SCIENCE

FOUNDATIONS OF PROBABILITY AND STATISTICS: APPLICAZIONE NEL MERCATO DELL'AUTOMOBILE

Relazione a cura di:

Alessandro Fossati, matricola 819499

Giorgio Nardi, matricola 819961

Anno Accademico 2019/2020

ABSTRACT

Il mercato dell'automobile è un mercato immenso e ricco di spunti in ambito statistico ed economico, per questo è risultato interessante rilevare un piccolo campione di vetture vendute tra il 2003 ed il 2018, il quale, seppur equivalendo circa ad un infinitesimo delle automobili vendute in questo range di tempo, permette di effettuare interessanti analisi di tipo statistico che permettono a loro volta di ottenere interessanti conclusioni di tipo economico. Conoscere come determinate caratteristiche di una automobile influenzano il suo prezzo di vendita risulta molto interessante ed utile, ed è proprio questo il fine ultimo della trattazione. Questo obiettivo diventa perseguibile con la formulazione di un modello lineare ed in particolare mediante la stima dei suoi coefficienti di regressione, i quali permettono di fornire interpretazioni e di fornire accurate giustificazioni riguardo la relazione tra il prezzo di vendita di automobile e le sue caratteristiche.

MATERIALI

Il dataset utilizzato proviene dalla piattaforma Kaggle. Esso è costituito da osservazioni riguardanti alcune automobili usate provenienti dal sito web cardekho.com. Il sito è di proprietà di Girnar Software, è stato lanciato nel 2008 e si occupa della vendita di auto usate. Il dataset è composto da 301 osservazioni e da 9 variabili:

- *Car_Name*: nome della macchina.
- *Year*: anno in cui è stata comprata la macchina.
- *Selling_Price*: prezzo di vendita della macchina.
- *Present_Price*: prezzo attuale della macchina nel concessionario in cui è stata acquistata.
- *Kms_Driven*: chilometri percorsi dalla macchina.
- *Fuel_Type*: tipo di carburante della macchina.
- *Seller_Type*: variabile dummy che definisce se chi vende la macchina è un rivenditore o un privato.
- *Transmission*: variabile dummy che definisce se la macchina è caratterizzata da cambio manuale o automatico.
- *Owner*: numero di possessori che la macchina ha avuto.

La variabile target, come anticipato nell'abstract, sarà *Selling_Price*.

PRE-PROCESSING

È utile osservare la tipologia delle variabili nella Tabella 1.

Tabella 1

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
1	Car_Name	0	0.00	0	0	0	0	factor	98
2	Year	0	0.00	0	0	0	0	integer	16
3	Selling_Price	0	0.00	0	0	0	0	numeric	156
4	Present_Price	0	0.00	0	0	0	0	numeric	147
5	Kms_Driven	0	0.00	0	0	0	0	integer	206
6	Fuel_Type	0	0.00	0	0	0	0	factor	3
7	Seller_Type	0	0.00	0	0	0	0	factor	2
8	Transmission	0	0.00	0	0	0	0	factor	2
9	Owner	290	96.35	0	0	0	0	integer	3

Non sono presenti valori mancanti all'interno del set di dati considerato. La tipologia della variabile *Transmission* non è adatta per l'analisi, si decide dunque di ricodificarla in una variabile factor. I livelli sono "prima", "seconda", "terza" e "quarta" che identificano se l'auto è di prima, seconda, terza o quarta mano. Inoltre, sono state create delle variabili aggiuntive utili per svolgere al meglio la fase di analisi descrittiva. Queste tuttavia non verranno incluse nel modello finale. Le nuove variabili sono:

- *Price_difference*: differenza tra *Present_Price e Selling_Price*. Essa è utile per avere un'idea della svalutazione dell'auto rispetto al mercato attuale.
- *Age*: differenza tra 2018 (anno a cui risalgono le rilevazioni) e la variabile *Year*. In questo modo si può avere un'idea di quanti anni è stata utilizzata una macchina.

ANALISI DESCRITTIVA

Si presentano nel Grafico 1 i boxplot riferiti alle variabili quantitative presenti nel set di dati. Tutte le distribuzioni sono caratterizzate da asimmetria positiva. Infatti, osservando i valori di media e mediana di ciascuna variabile si riscontra che il primo indice risulta sempre essere maggiore del secondo. Un altro tratto comune all'interno dei boxplot è la presenza di outliers, ossia di valori che superano il terzo quartile di almeno 1.5 volte la distanza interquartilica. Bisogna considerare che queste osservazioni influenzano simmetria e media delle distribuzioni.

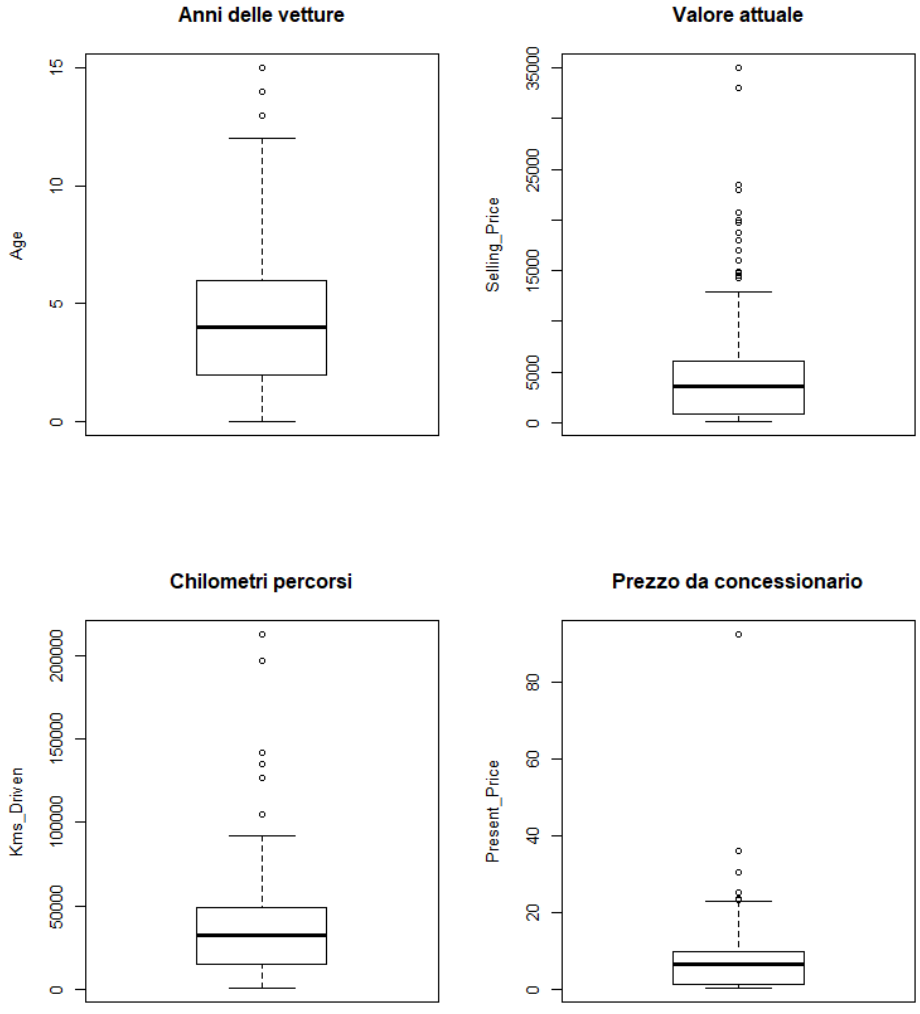

Grafico 1

Negli istogrammi (Grafico 2) viene confermato quanto visto all'interno dei boxplot (Grafico 1). Infatti, essi mostrano in ciascuna variabile una asimmetria positiva evidenziata dalla presenza di una gobba spostata verso sinistra. Si può dunque affermare che nessuna variabile segua distribuzione Normale.

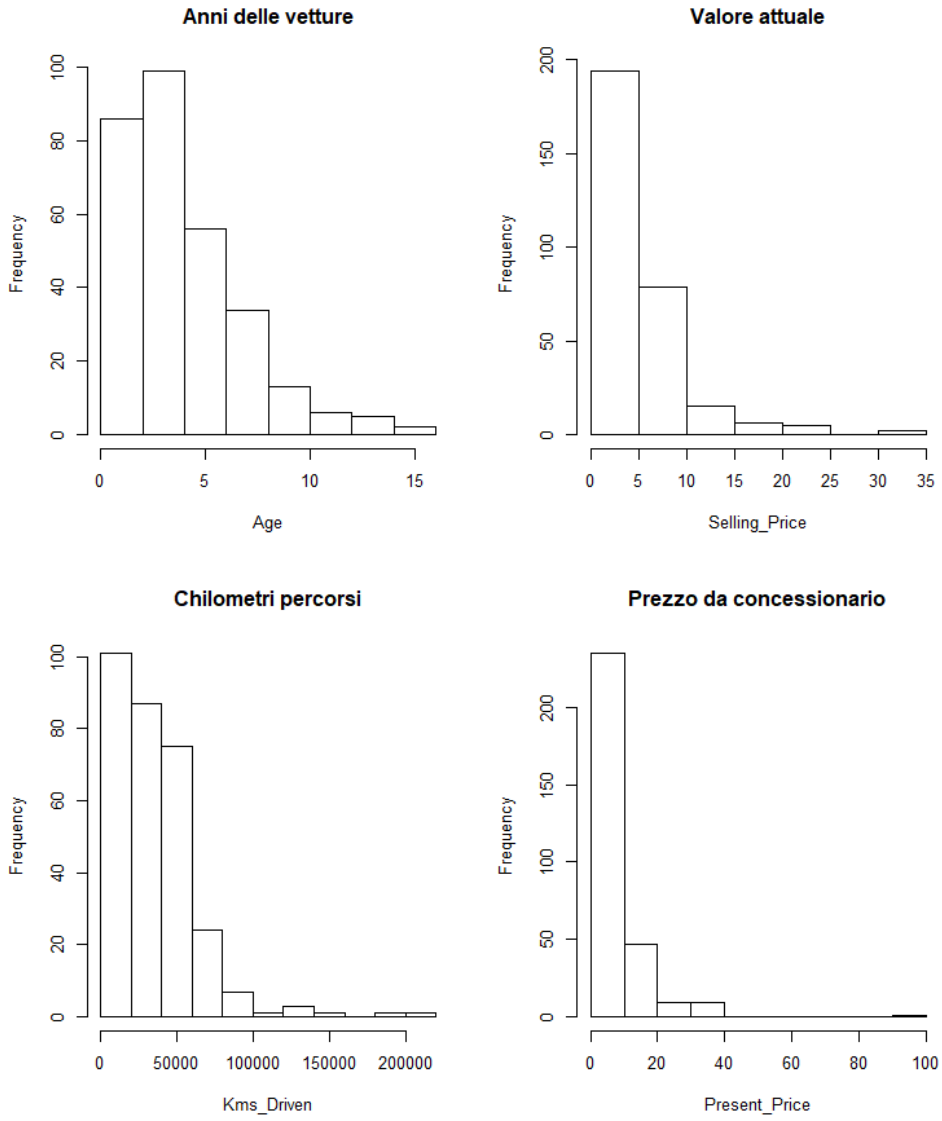

Grafico 2

Per confrontare la variabilità degli attributi quantitativi viene calcolato il coefficiente di variazione di ciascuno di essi, come rapporto tra deviazione standard e valore assoluto della media, in modo tale da annullare l'effetto dell'unità di misura. l'attributo più variabile secondo tale indice risulta essere *Age* (cv=1.512), mentre quello meno variabile è *Present_Price* (cv=0.883).

Proseguendo con l'analisi descrittiva, si è voluto studiare il livello di connessione tra le variabili *Transmission e Age*. Infatti, siccome la vendita di auto con cambio automatico si è sviluppata soprattutto negli ultimi anni è possibile che il campione considerato rifletta tale fenomeno. Per fare ciò è stato necessario ricodificare la variabile Age in una factor di livelli: "età bassa" (meno di 5 anni), "età media" (più di 5 anni e meno di 10), "età alta" (più di 10 anni). Inoltre, è utile valutare la composizione della popolazione in base a cambio automatico o manuale con il diagramma a settori circolari del Grafico 3.

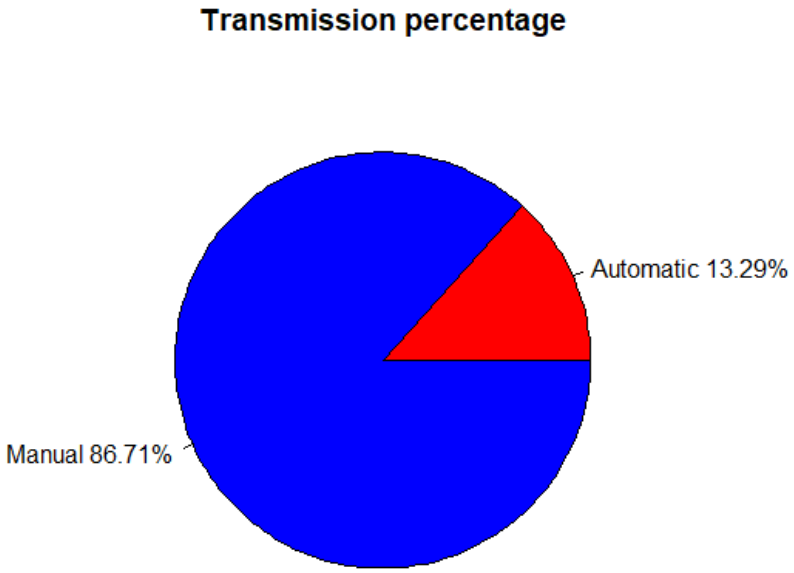

Grafico 3

Osservando la sintesi tabellare delle due variabili rilevate congiuntamente con le frequenze percentuali di colonna si nota che Il 65% delle auto con cambio automatico che costituiscono il campione ha un'età bassa. Tuttavia, bisogna considerare che le auto con età bassa costituiscono gran parte del campione, ovvero il 61,4% del totale. Per misurare la connessione tra le 2 variabili è stato calcolato l'indice Chi quadro di Pearson normalizzato, che risulta essere uguale a 0,0013. Esso è prossimo a 0 quindi non sembra esserci alcuna connessione tra Age e Transmission.

Il grafico a barre (Grafico 4) riportato nella pagina seguente, mostra che automobili con cambio manuale o automatico si distribuiscono in proporzione molto simile tra le 3 classi di età, di conseguenza esso conferma che non vi è alcuna connessione tra le due variabili categoriali:

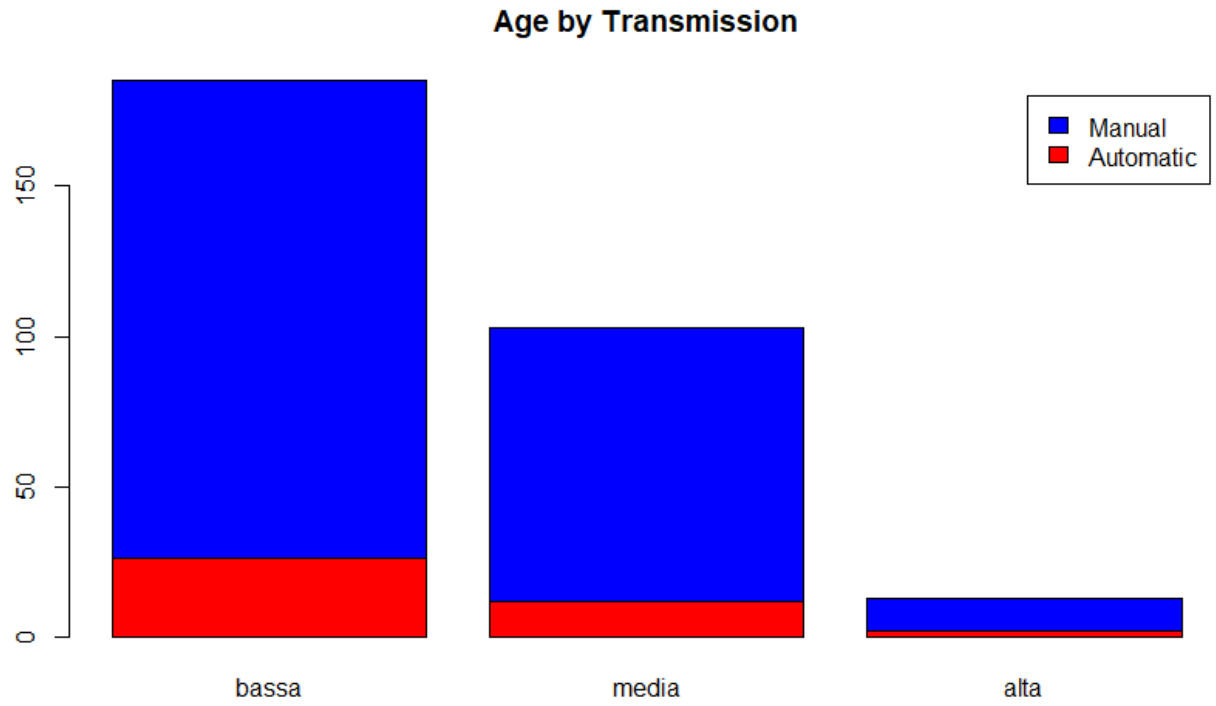

Grafico 4

Inoltre, si è deciso di verificare graficamente l'eventuale presenza di correlazione tra le variabili *Price_difference e Kms_Driven*. Infatti, si può credere che più una macchina usata ha percorso strada più questa di conseguenza si svaluterà. Il Grafico 5 rappresenta quindi uno scatterplot bivariato condizionato per la variabile *Age*.

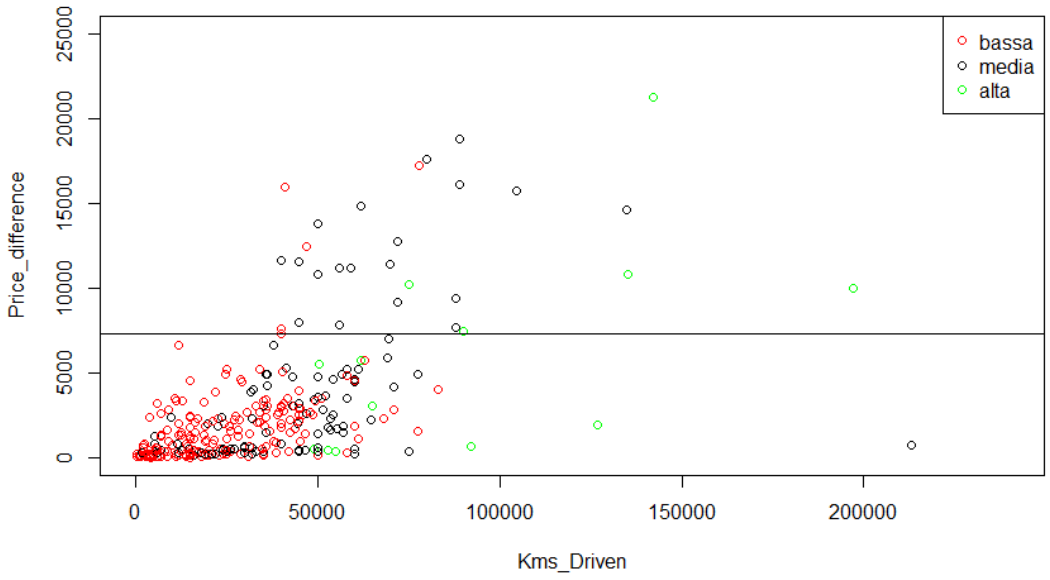

Grafico 5

Quest'ultimo (Grafico 5) identifica la presenza di una correlazione positiva (0.49) tra la differenza di prezzo e i chilometri percorsi. Inoltre, si può dedurre, come è logico che sia, che le auto che hanno percorso più strada sono caratterizzate da un'età media o alta. Il fatto che gran parte delle auto che mostra una svalutazione al di sopra del nono decile della distribuzione, ovvero 7350$ di svalutazione, ha un'età media o alta fa dedurre che l'età dell'auto è una variabile discriminante riguardo la svalutazione. Quindi, si può concludere che la quantità di chilometri percorsi influenza certamente la differenza di prezzo, ma anche la variabile *Age* ha un impatto da non trascurare. Osservando attentamente, nasce il sospetto che nelle auto di età media la correlazione tra *Price_difference e Kms_Driven* sia maggiore rispetto alle altre 2 classi di età. A questo proposito è utile osservare gli scatterplot univariati per ogni classe di età (Grafico 6).

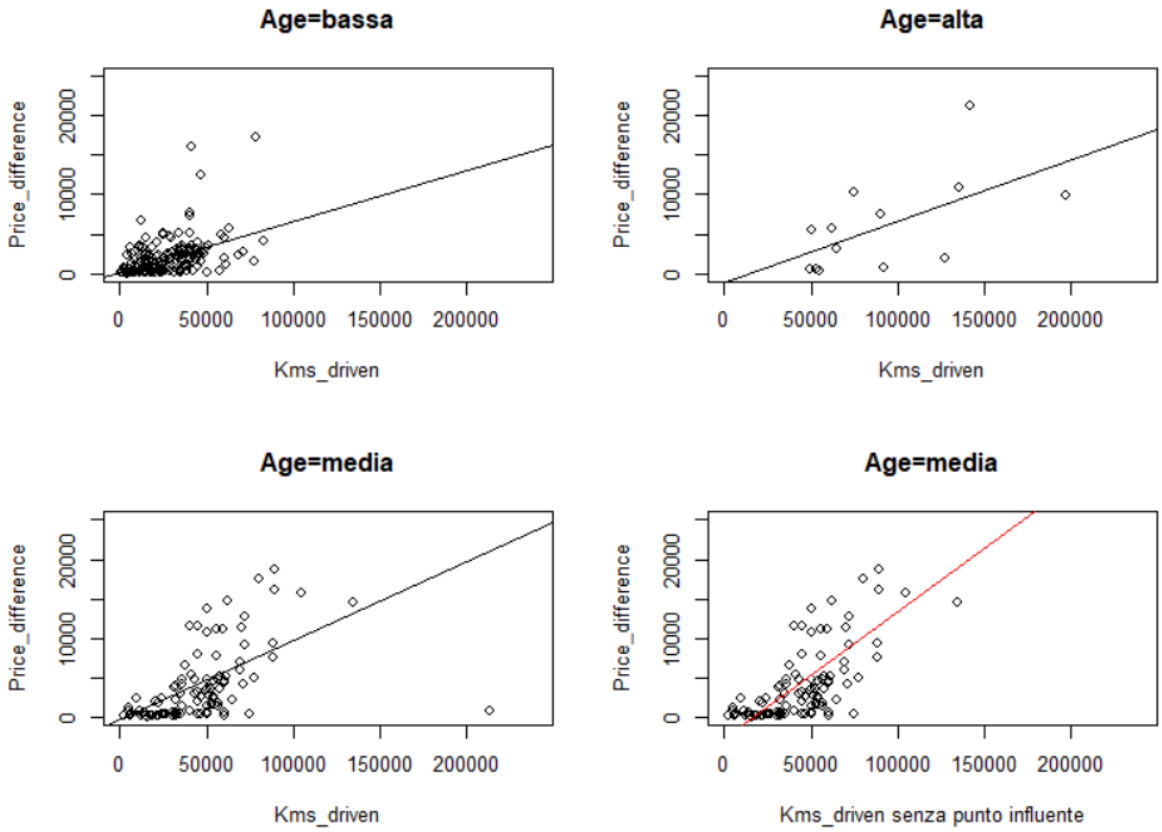

Grafico 6

Inizialmente, confrontando i modelli di regressione lineare semplice relativi alle classi di età bassa e media, si osserva in entrambi i casi un'elevata significatività relativa al parametro associato all'unica covariata presente nei modelli. Inoltre, il modello elaborato sulle auto di età bassa presenta un valore del coefficiente di determinazione lineare (R^2=0.23) maggiore rispetto a quello del modello stimato sulle vetture con età media (R^2=0.17). Tuttavia, nel grafico relativo alle auto di età media (Grafico 6c), si nota la presenza di un punto che potrebbe essere molto influente sulla stima del modello, ossia quella osservazione che supera i 200000 km percorsi. Di conseguenza, escludendo tale osservazione dal modello, esso ne trae beneficio in termini di significatività dei parametri e

bontà di adattamento ai dati(R^2=0.3). Considerando la regressione relativa alle auto di età alta, il modello risulta essere migliore in termini di R^2 rispetto agli altri modelli considerati. Tuttavia, il modello è fortemente condizionato dalle poche osservazioni ad esso inerenti. In aggiunta, il coefficiente relativo a *Kms_Driven* risulta essere meno significativo rispetto ai coefficienti stimati negli altri modelli, dove per significatività si fa riferimento al test di Wald, per il quale si è indotti a rifiutare l'ipotesi nulla di parametro β uguale a zero in tutti i modelli stimati precedentemente.

Procedendo con l'analisi bivariata, potrebbe essere interessante analizzare la distribuzione del prezzo da concessionario al variare della tipologia di carburante (Grafico 7).

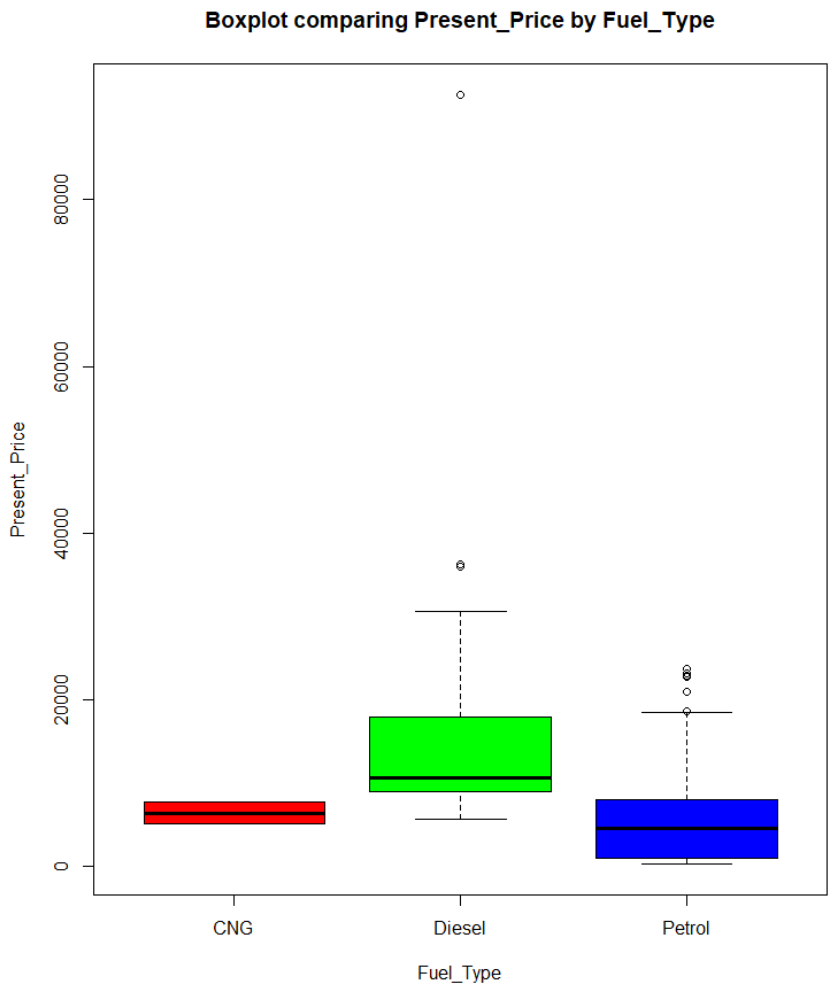

Grafico 7

La distribuzione rispetto a CNG risulta difficilmente interpretabile in quanto vi sono solo 2 osservazioni ad essa inerenti nel set di dati. La distribuzione rispetto a Diesel risulta essere asimmetrica positivamente poiché la sua media (15815$) è maggiore della sua mediana (10585$). Bisogna inoltre sottolineare che sono presenti alcuni outliers, i quali indubbiamente influenzano la simmetria e il valore della media. Anche per quanto riguarda la distribuzione del prezzo rispetto a

Petrol si evince asimmetria positiva, in quanto la media (5584$) risulta essere maggiore della mediana(4600$). In aggiunta, si può notare che quest'ultima presenta una minore asimmetria rispetto alla distribuzione riguardante le auto Diesel.

STIMA E TEST D'IPOTESI

Come primo passo, si decide di calcolare la stima intervallare per la variabile target *Selling_Price*. L'obiettivo è quello di ottenere dalle osservazioni campionarie un insieme di valori plausibili per l'ignoto parametro media della popolazione. Si potrebbe stimare la media della popolazione mediante lo stimatore media campionaria. Tuttavia, trattandosi di una stima puntuale, non vi è possibilità che la stima sia esattamente uguale alla media. Di conseguenza, risulta necessario usare la stima di tipo intervallare. Si stabilisce un livello di confidenza del 95% e si assume che la popolazione abbia distribuzione normale con varianza ignota. Dunque, si utilizza lo stimatore varianza campionaria per ottenere una stima della varianza della popolazione. l'intervallo risulta essere: (4087,5236)$. In generale, l'intervallo di confidenza con grado di fiducia al 95% va interpretato nel modo seguente: considerando tutti i possibili campioni di ampiezza n=301 auto usate e calcolando per ciascuno di essi la media campionaria relativa al prezzo di vendita ed il corrispondente intervallo di confidenza su di essa centrato, il 95% degli intervalli così ottenuti conterrà il prezzo medio di vendita della popolazione (μ), mentre il 5% non lo conterrà.

Si vuole ora stimare il valore del parametro *p (probabilità di successo)*, che rappresenta la frequenza relativa o proporzione con cui una vettura viene venduta da un privato. Si considera dunque un campione di ampiezza n=301 dalla popolazione e sia la proporzione campionaria *P*(stimato)=106/301, dove 106 è il numero di volte in cui un'auto viene venduta da un privato nel campione di riferimento. La distribuzione binomiale, di parametri *n=301 e p*, può essere approssimata da una distribuzione normale avente $\mu=np$ *e* $\sigma^2=np(1-p)$. Si stabilisce un grado di fiducia del 95%. La stima intervallare ottenuta risulta essere: (0.30,0.41). Nuovamente, l'intervallo con livello di confidenza al 95% va interpretato come segue: considerando tutti i possibili campioni di ampiezza n=301 auto usate e calcolando per ciascuno di essi la proporzione campionaria relativa alla tipologia di venditore privato ed il corrispondente intervallo di confidenza su di essa centrato, il 95% degli intervalli così ottenuti contiene la proporzione di venditori privati della popolazione(p) e solo il 5% non lo contiene.

Risulta inoltre interessante effettuare alcuni test d'ipotesi. ACI (Automobile Club d'Italia) studio e Ricerche ha svolto un'indagine di mercato nel 2016 secondo cui è emerso un dato significativo: Il 74% degli intervistati non prenderebbe in considerazione l'acquisto di veicoli usati con più di 90000 chilometri. Si vuole dunque testare l'ipotesi nulla per cui il numero medio di chilometri percorsi dalle automobili della popolazione sia maggiore o uguale a 90000. In modo tale da valutare se le vetture della popolazione sono appetibili sul mercato in base al campione considerato. Si fissa a priori un livello di confidenza del 95%.

$$H_0: \mu \geq 90000$$

$$H_1: \mu < 90000$$

I dati considerati provengono da una popolazione di cui non si conosce la distribuzione. La varianza è ignota, dunque, verrà utilizzato lo stimatore varianza campionaria. In aggiunta, Il campione preso in esame ha un'ampiezza n grande (molto maggiore di 30). Di conseguenza, la statistica test assume la seguente forma:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$

Il valore osservato della statistica test (Z_{oss}=-33.55) cade nella regione di rifiuto, la quale comprende i valori inferiori al quantile negativo della normale standard considerando un livello di confidenza pari a 0.95 ($Z_{0.05}$=1.64). Dunque, si può affermare che, data l'evidenza campionaria, si può rifiutare l'ipotesi nulla che il numero medio dei chilometri percorsi dalle auto usate che compongono la popolazione sia maggiore di 90000, con un livello di significatività (α) del 5%. Di conseguenza, sembra che le vetture usate facenti parte della popolazione possano rientrare negli interessi dei consumatori.

In seguito, si decide di testare, utilizzando il campione, se la media del prezzo di vendita delle auto con cambio automatico possa essere ritenuta uguale a quella delle vetture con cambio manuale. Questa volta non verrà fissato nessun livello di significatività a priori.

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Si assume che le osservazioni dei 2 blocchi provengano da due popolazioni di distribuzione normale. Le varianze delle due popolazioni sono ignote, quindi, viene stimata la *varianza pooled*, dopo aver effettuato il calcolo delle varianze campionarie relative a ciascun blocco di osservazioni.

La statistica Test sotto H_0 si distribuisce come una T-student con 41 gradi di libertà e con valore osservato pari a -3.91. Il p-value, che indica la probabilità di ottenere un risultato uguale o più estremo di quello osservato supposta vera H_0, equivale a 0.0003. Di conseguenza, si rifiuta l'ipotesi nulla persino ad un livello di significatività dell'1%. Dunque, si può affermare che in base all'evidenza campionaria le auto usate con cambio manuale presentano media di prezzo differente rispetto a quella relativa alle automobili con cambio automatico.

Infine, si effettua un test ANOVA per testare se al variare della tipologia di carburante consumato almeno una delle medie di prezzo di vendita risulta essere diversa dalle altre. L'ipotesi alla base consiste nel fatto che le osservazioni di ciascun blocco provengano da popolazioni che seguono una distribuzione normale con uguale varianza e con eventuale media diversa per alcune popolazioni.

$$H_0: \mu_1=\mu_2=\mu_3$$

$$H_1: \text{almeno una media diversa}$$

Sotto H_0 la statistica test si distribuisce come una Fisher-Snedecor con k-1=2 e n-k=298 gradi di libertà. Il p-value è prossimo a 0, dunque, si rifiuta l'ipotesi nulla di uguaglianza tra le medie di ciascun gruppo a qualsiasi soglia classica di significatività.

ELABORAZIONE DEL MODELLO

In primo luogo, si vuole effettuare un'analisi preliminare sulla matrice di correlazione (Grafico 8) tra le variabili esplicative di tipo quantitativo per trarne informazioni sulla possibile presenza di multicollinearità. Quest'ultima è implicata dal fatto che le variabili esplicative risultano altamente correlate fra di loro. Di conseguenza, in tale situazione, i coefficienti di regressione risulterebbero probabilmente instabili e le statistiche T per le variabili risulterebbero errate.

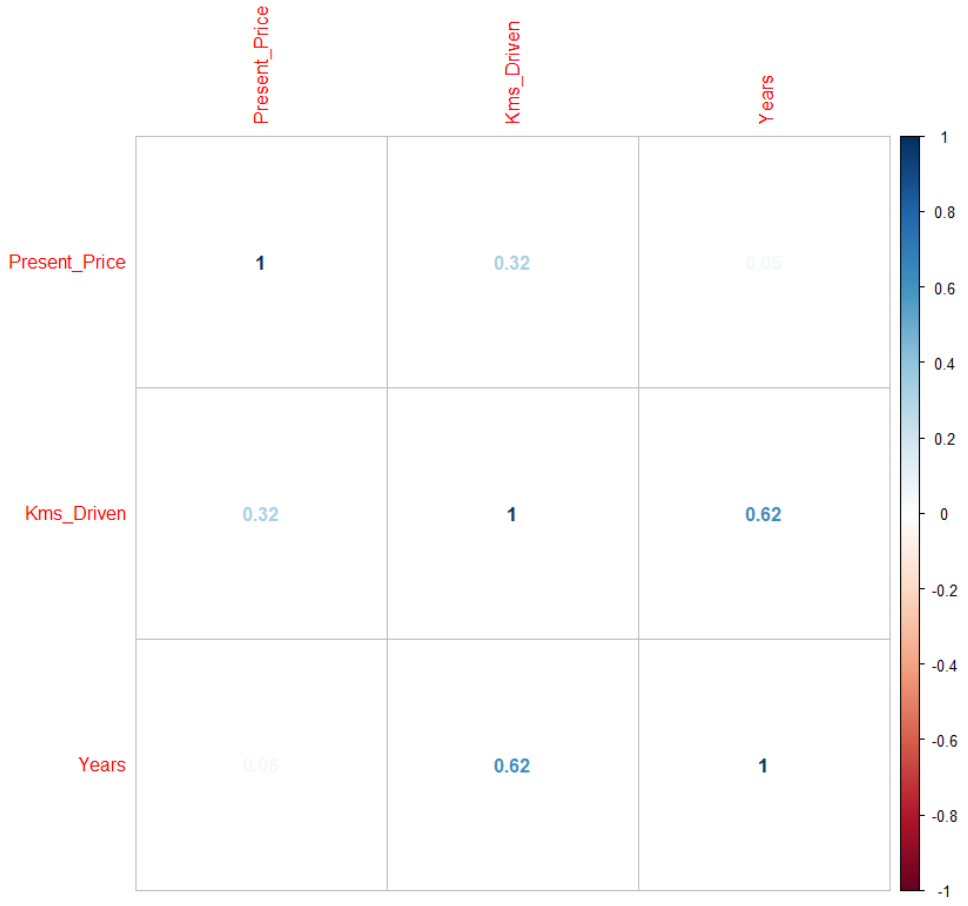

Grafico 8

Fissato un valore soglia pari a 0.85, superato il quale, due variabili risulterebbero troppo correlate fra loro, la matrice di correlazione non mostra alcun valore preoccupante in questo senso. Tuttavia, per avere la certezza dell'assenza di multicollinearità verrà effettuata l'analisi dei VIF (variance inflation factor) dopo la stima del modello.

Considerando come variabile target *Selling_Price*, si stima il seguente modello di regressione lineare multipla con il metodo OLS (Ordinary Least Squares):

$$\mu(Selling\_Price_i) = 4100.5 + 0.45(Present\_Price_i) - 0.019(Kms\_Driven_i) + 2498.7(Fuel\_TypeDiesel_i) + 502.1(Fuel\_TypePetrol_i) - 1173.8(Seller\_TypeIndividual_i) - 1440.4(TransmissionManual_i) + 366.7(OwnerSeconda_i) - 4969.5(OwnerQuarta_i) - 323.3(Years_i)$$

Il test F, che si utilizza per verificare se esiste una relazione significativa tra la variabile dipendente e l'insieme delle variabili indipendenti, induce a rifiutare l'ipotesi nulla $H_0: \beta_1 = \beta_2 = \ldots \beta_9 = 0$, in quanto il p-value risulta prossimo a 0. Inoltre, i valori di R^2 ed R^2-adjusted risultano molto alti (rispettivamente 0.89 e 0.887), dunque, anche la devianza spiegata risulta essere elevata. Si evincono infine valori alti di significatività per quasi tutti i parametri del modello. Tutte queste caratteristiche inducono ad affermare che il modello potrebbe adattarsi in maniera soddisfacente ai dati.

Si fornisce di seguito qualche interpretazione relativa ai parametri del modello:

- A parità di tutte le altre covariate, se l'età di un'automobile aumenta di un anno, il suo prezzo di vendita diminuisce di 323.3$ in media.
- A parità di tutte le altre covariate, passando dalla categoria di auto con cambio automatico a quella con cambio manuale, il prezzo di vendita diminuisce di 1440.4$ in media.
- A parità di tutte le altre covariate, passando dalla categoria di auto di prima mano a quella di seconda mano, il prezzo di vendita aumenta di 366.7$ in media. Tuttavia, questo parametro non risulta essere significativo rispetto alle soglie classiche di significatività. Infatti, logicamente, non sembra rispecchiare ciò che avviene nella realtà dei fatti.

In aggiunta, analizzando i valori ottenuti mediante il calcolo dei VIF per ciascuna variabile esplicativa, si riscontra definitivamente l'assenza di multicollinearità. Infatti, la radice quadrata di ognuno di questi valori risulta essere minore del valore soglia 2.

Come ultima analisi ci si concentra sui residui, in particolare si vuole verificare l'ipotesi di normalità dei residui, assunzione fondamentale per la costruzione di un buon modello lineare. Per fare ciò si presenta il grafico a pagina seguente (Grafico 9):

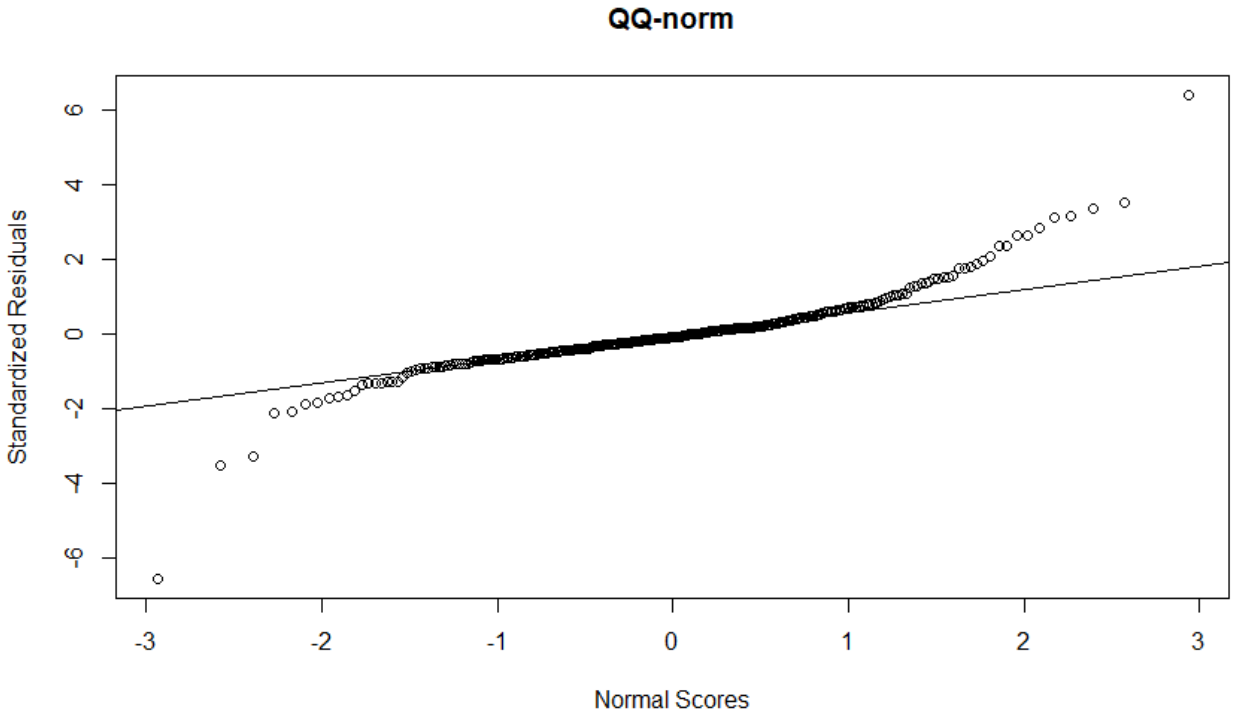

Grafico 9

Il normal QQ plot è un grafico che confronta i valori dei residui standardizzati con la linea che individua la loro distribuzione normale, ovvero il grafico rappresenta una figura per cui se i punti si distribuiscono sulla linea la distribuzione dei residui risulta normale. L'ipotesi di normalità non è del tutto rispettata in quanto le code della distribuzione si discostano dai quantili teorici. Risulterebbero dunque necessarie ulteriori analisi inerenti ai residui, le quali non verranno approfondite in questa trattazione, in quanto non conformi agli obiettivi prefissati