Analisi dei prezzi delle automobili usate

Advanced Statistical Modelling For Big Data
Andrea Mauro [mat.0222400914]

Introduzione

Nel seguente progetto viene preso in analisi il fenomeno dei prezzi delle automobili usate in India, in particolare, il principale obiettivo del lavoro proposto è la ricerca dei fattori dominanti cui determinano il prezzo dell'automobile usata, dunque in questo contesto saranno utilizzati strumenti di analisi esplorativa per catturare le peculiarità del fenomeno. Inoltre sarà proposta una sezione di modellistica, dove si cercerà di stimare un modello in grado di cogliere e riprodurre le dinamche del fenomeno di interesse: i prezzi. Nell'ultima parte del lavoro saranno quindi anche valutate le capacità previsive dei modelli stimati e forniti degli esempi circa alcuni dei fattori determinanti nel fenomeno dei prezzi delle autovetture usate.

1 Dati

Il dataset preso in esame si compone di 6019 osservazioni per ciascuna delle quali sono registrate 12 caratteristiche relative all'autovettura, in particolare, le informazioni riguardano: il modello dell'autovettura, la casa produttrice, il prezzo, l'anno di produzione, il luogo nel quale l'auto è messa in vendita, il numero di posti a sedere, tipologia del motore, la cilindrata, i cavalli, la trasmissione che monta l'automobile, l'efficienza del carburante , il chilometraggio e il numero di precedenti possessori.

Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price	company
2010	72000	CNG	Manual	First	26.60	998	58.16	5	1.75	Maruti
2011	46000	Petrol	Manual	First	18.20	1199	88.70	5	4.50	Honda
2013	40670	Diesel	Automatic	Second	15.20	1968	140.80	5	17.74	Audi
2012	75000	LPG	Manual	First	21.10	814	55.20	5	2.35	Hyundai
2013	86999	Diesel	Manual	First	23.08	1461	63.10	5	3.50	Nissan
2016	36000	Diesel	Automatic	First	11.36	2755	171.50	8	17.50	Toyota

Figure 1: Prime 6 osservazioni del set di dati

Prima di procedere con le analisi delle caratteristiche del set di dati, la fase di *Data Wrangling* si rende necessaria in quanto il set originale si presenta in forma grezza, dunque sono state formattate le tipologie delle variabili presenti nel campione, in particolare la formattazione utilizzata è la seguente:

- *Year*: variabile numerica, rappresenta l'anno di produzione dell'automobile;
- *Kilometers_Driven*: variabile numerica, rappresenta il chilometraggio dell'automobile;
- *Fuel_Type*: variabile categorica nominale, rappresenta l'alimentazione del motore (presenti 4 categorie: GPL, Benzina, Diesel, Cng);
- *Transmission*: variabile categorica nominale, rappresenta la trasmissione del veicolo (presenti 2 categorie: Automatico e Manuale);
- *Owner_Type*: variabile categorica ordinale, rappresenta il numero di precedenti possessori del veicolo (presenti 3 categorie: Uno, Due, Tre o Più, con Uno>Due>Tre o più);
- *Mileage*: variabile numerica, rappresenta l'efficienza del veicolo in termini di carburante, misurata attraverso i chilometri percorsi con un litro (km/l);
- *Engine*: variabile numerica, rappresenta la cilindrata del motore, espressa in cm^3;

- *Power*:variabile numerica, rappresenta il numero dei cavalli vapore del motore (CV);
- *Seats*: variabile categorica nominale, rappresenta il numero di posti a sedere del veicolo (presenti 6 categorie: Due,Quattro,Cinque,Sei,Sette,Otto);;
- *Price*: variabile numerica, rappresenta il prezzo al quale l'automobile viene venduta, si specifica che un'unità di tale variabile rappresenta 1000 dollari statunitensi;
- *Company*: variabile categorica nominale, rappresenta la casa produttrice dell'automobile (presenti 31 categorie).

In seguito alla formattazione delle feature, vengono rimosse le osservazioni ridondanti e gli oggetti che presentano alcune caratteristiche non disponibili. Dunque in seguito alla fase di *handling* del set di dati, il campione risulta essere composto da 5836 *osservazioni*, per cui una riduzione del 3% del campione iniziale.

2 Analisi Esplorativa

In questa fase vengono approfondite le caratteristiche della varibile *Prezzo*, analizzandone la struttura e le relazioni con altre feature presenti nel campione. Chiaramente questa fase è fondamentale per comprendere le dinamiche del prezzo delle automobili usate e fornire ottimi spunti informativi da cui partire per la successiva modellazione del fenomeno di interesse attraverso gli strumenti presenti nella letteratura statistica.
Dunque il primo passo per comprendere il fenomeno del Prezzo delle automobili usate è quello di osservare ed analizzare la struttura della variabile *Prezzo* presente nel campione. Di seguito viene mostrata la figura rappresentante la funzione di densità della feature *Prezzo*:

Figure 2: Densità della variabile *Prezzo*

Innanzitutto, considerando la variabile *Prezzo*, per sua natura, questa si manifesta attraverso valori non negativi in modo continuo. Dal grafico si osserva chiaramente come la distribuzione del *Prezzo* risulti chiaramente asimmetrica positiva, caratterisitica evidenziata dalla lunga coda destra della distribuzione (il cui valore estremo è pari a 150 mila dollari), vengono mostrati nella figura anche il valore mediano e medio della distribuzione. La mediana fornisce un'utile prospettiva sulle autovetture presenti nel campione, in particolare, il fatto che la mediana sia pari a 5.75 suggerisce che la metà delle osservazioni ha un prezzo inferiore a questo valore.

La Figura 3 mette in evidenza la densità della variabile *Prezzo* condizionata per il tipo di *Trasmissione* che monta l'autovettura:

Figure 3: Densità della variabile *Prezzo* condizionata per la *Trasmissione*

Le due densità condividono una simile asimmetria positiva, la differenza principale emerge nelle code: in particolare, la coda destra della densità dei prezzi delle automobili con cambio automatico risulta maggiore rispetto a quella delle automobili con trasmissione manuale.
Successivamente, viene presa in esame la distribuzione condizionata del prezzo in relazione alla tecnologia utilizzata per alimentare il motore, attraverso l'utilizzo dello strumento grafico del Box-Plot:

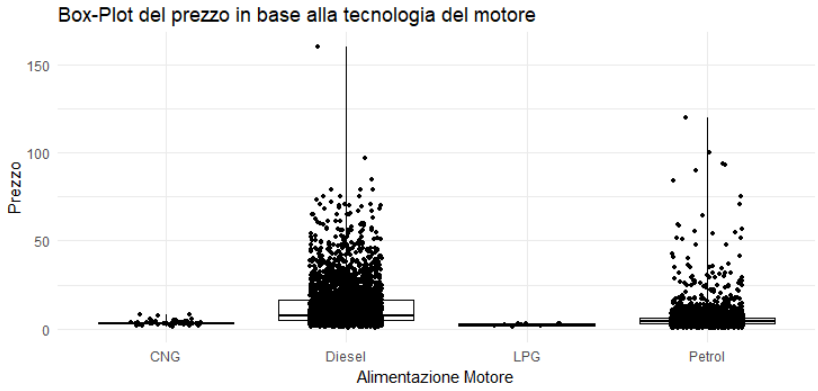

Figure 4: Box-plot Prezzo condizionato per tipo di alimentazione motore

Inanzitutto, osservando la Figura 4 è netto lo sblianciamento delle classi per la variabile *Fuel_Type*, in particolare, nel campione sono presenti poche centinaia di autovetture il cui motore è alimentato a Gas Naturale Compresso(CNG) e Gas di Petrolio Liquefatto (LPG). La distribuzione dei prezzi delle autovetture alimentate a gas naturale compresso e gas di petrolio liquefatto appare sostanzialmente simile, con valori mediani di circa 3.25 mila dollari per il CNG e 2.6 mila dollari per il GPL.
Nel confrontare le due tipologie di carburante "popolari" nel campione, ossia la benzina e il diesel, emerge un'interessante evidenza: il primo quartile della distribuzione dei prezzi delle automobili con motore alimentato a diesel presenta un prezzo massimo maggiore della metà delle autovetture alimentate a benzina.

Un simile approccio è stato impiegato per esplorare le peculiarità della distribuzione della variabile prezzo condizionata per il numero di precedenti possessori dell'automobile, contrariamente alle aspettative, non sono emerse evidenze significative in questo contesto.

Proseguendo con l'analisi esplorativa del campione preso in esame, di seguito vengono approfondite struttura e peculitarità della variabile *Prezzo* in relazione ad alcune variabili di tipo numeriche presenti nella raccolta dati.

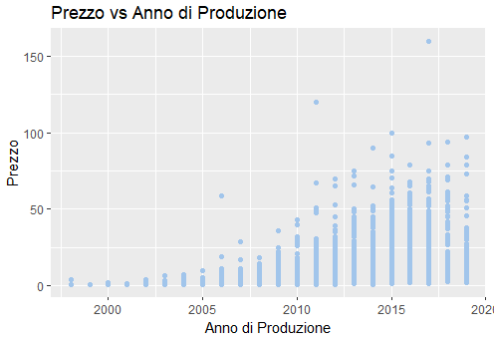

Figure 5: Scatterplot *Prezzo-Anno Prod.*

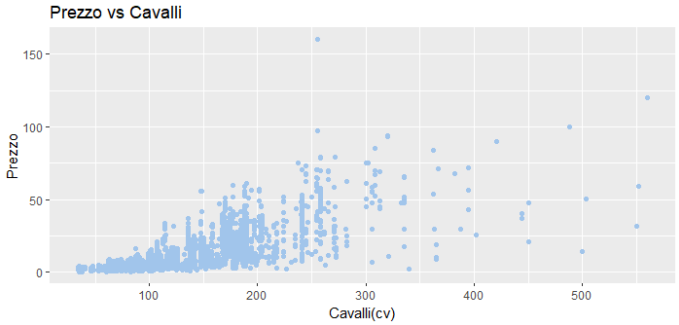

Figure 6: Scatterplot *Prezzo-Cavalli*

Attraverso l'impiego dello strumento grafico a dispersione, siamo in grado di cogliere struttura e relazione che intercorre tra la variabile di interesse prezzo e l'anno di produzione dell'autovettura (Figura 5), dalla figura si nota chiaramente come la varianza del prezzo incrementi con l'aumentare del livello dell'anno di produzione.
Ugualmente, osservando la Figura 6 notiamo come la stessa peculiarità si riproponga nella relazione tra il prezzo dell'automobile e i suoi Cavalli.

Continuando con l'analisi delle relazioni presenti tra la variabile Prezzo ed altre feature numeriche:

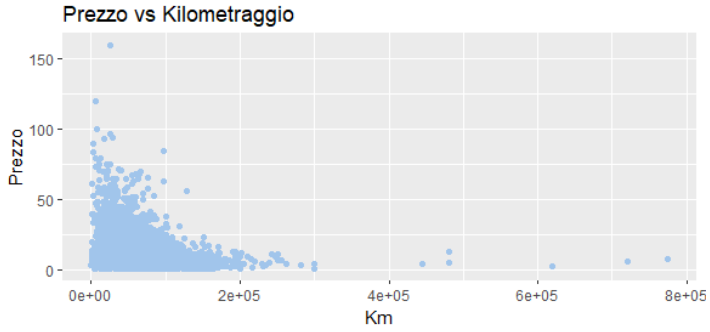

Figure 7: Scatterplot *Prezzo-km*

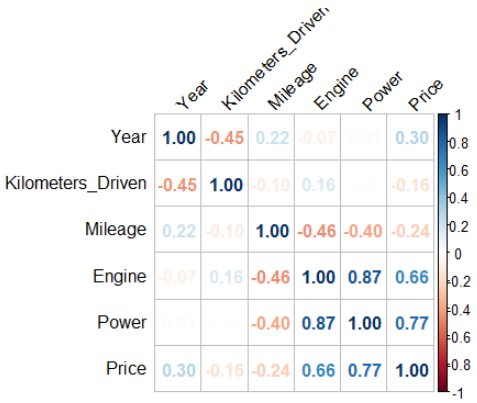

Figure 8: Matrice delle Correlazioni

Dalla Figura 7 emerge la relazione esistente tra il *kilometraggio* e il *prezzo* dell'automobile. In particolare, si osserva che all'aumentare del kilometraggio, il prezzo tende a diminuire, indicando una presunta relazione(non necessariamente lienare) negativa tra le due variabili.
Attraverso la matrice delle correlazioni (Figura 8), viene fornita una misura dell'intensità della relazione lineare che intercorre tra le variabili. Per quanto riguarda la variabile *Prezzo* osserviamo come questa presenti relazioni lineari d'intensità piuttosto moderata con le variabili *Cilindrata* (Engine) e *Cavalli* (Power). Per quanto riguarda altre relazione lineari, si evince una forte intensità del legame tra le feature *Cilindrata* e *Cavalli*.

3 Modello

In seguito all'approfondimento condotto durante la fase di analisi esplorativa, volta a esaminare dettagliatamente la struttura e le peculiarità del fenomeno di interesse, emerge chiaramente che l'insieme dei valori assunti dalla variabile *Prezzo* appartiene al dominio dei valori continui non negativi. Inoltre, si nota che la distribuzione di tali valori manifesta un carattere asimmetrico positivo.

Date le caratteristiche emerse, in un contesto di modellazione del fenomeno dei Prezzi si può assumere che il medesimo segua approsimativamente una distribuzione *Gamma* o *Gaussiana Inversa*, queste distribuzioni risultano essere nautrali candidate per modellare il *Prezzo* in quanto presentano asimmetria positiva (nel dominio dei valori positivi e continui), con un picco netto ed una lunga coda a destra. Difatti le due distribuzioni presentano prorietà simili, la *Gaussiana Inversa* è caratterizzata da maggiore asimmetria e un picco più netto.
Dunque nella fase di modellazione, vengono prese in considerazione due regressioni distinte. In una di esse, si ipotizza che la variabile dipendente condizionata sui regressori (*Prezzo*) segua una distribuzione *Gamma*, mentre nell'altra si assume una distribuzione *Gaussiana Inversa*.
Dopo aver convertito le diverse *variabili categoriali* presenti nel dataset in *variabili dummy*, il totale dei regressori disponibili nella fase di modellazione è risultato essere 44. Per la selezione delle variabili da incorporare nella regressione, abbiamo inizialmente adottato un approccio di stima del modello che comprendeva tutti i regressori a disposizione. Successivamente, abbiamo escluso le variabili i cui coefficienti non raggiungevano un livello di significatività pari a 0,05 (ossia, p-value $> 0{,}05$), generando così il modello ridotto.

Regressoione Gamma
La regressione Gamma risulta essere una particolare specificazione del *modello lineare genralizzato* nel quale si assume che distribuzione della componente casuale sia Gamma.
In questo contesto la *funzione logaritmica* è stata scelta per collegare la media e i regressori.
Il modello completo include tutte le variabili, mentre per il modello ridotto ne sono escluse 10.

```
(Dispersion parameter for Gamma family taken to be 0.06102339)

    Null deviance: 3568.37  on 4375  degrees of freedom
Residual deviance:  265.38  on 4332  degrees of freedom
AIC: 16437
```

Figure 9: Modello completo

```
(Dispersion parameter for Gamma family taken to be 0.06093331)

    Null deviance: 3568.37  on 4375  degrees of freedom
Residual deviance:  266.35  on 4342  degrees of freedom
AIC: 16433
```

Figure 10: Modello ridotto

La discriminante utilizzata per la selezione del modello finale viene fornita dalla bontà di adattamento di ciascun modello. Evidenza particolare viene fornita dal valore della devianza residuale per i due modelli, in particolare i valori differiscono per poco meno di 1, per cui l'apporto informativo alla log-verosimiglianza delle 10 variabili escluse risulta approssimativamente nullo. Ugualmente simili performance si registrano con il criterio di informazione Akaike.
In generale, la bontà di adattamento dei due modelli risulta essere complessivamente simile per cui la discriminante applicata è la parsimonia dei modelli, per tale motivo viene preferito quello ridotto. Chiaramente la selezione del modello parsimonioso viene validata attraverso il Likelihood Ratio Test, se i due modelli non risultassero equivalenti non potremmo ignorare le differenze esistenti tra i due modelli.

Abbiamo quindi confrontato tramite il Likelihood Ratio Test (LRT) il modello completo, quello ridotto e quello con il solo termine dell'intercetta.

```
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
     4342     266.35
     4332     265.38 10  0.96571   0.1047
```

Figure 11: Test LRT, confronto modello ridotto e completo

```
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
     4375     496.19
     4342     266.35 33   229.84 < 2.2e-16 ***
```

Figure 12: Test LRT, confronto modello ridotto e modello intercetta

Secondo il Likelihood Ratio Test, non rifiutiamo l'ipotesi nulla di equivalenza tra il modello completo e quello ridotto (riferimento Figura 11), mentre rifiutiamo l'equivalenza tra il modello che include solo

l'intercetta e il ridotto (riferimento Figura 12). Dunque viene così confermata la scelta finale di usare il modello ridotto per il principio di parsimonia.

Regressione Gaussiana Inversa

La regressione Gaussiana Inversa risulta essere un'altra particolare specificazione del *modello lineare generalizzato* nel quale si assume che distribuzione della componente casuale sia Guassiana Inversa.
In questo contesto la *funzione logaritmica* è stata scelta per collegare la media con i predittori.
Il modello completo include tutte le variabili, mentre per il modello ridotto ne sono escluse 11.

```
(Dispersion parameter for inverse.gaussian family taken to be 0.01276154)

    Null deviance: 496.194  on 4375  degrees of freedom
Residual deviance:  58.273  on 4332  degrees of freedom
AIC: 17833
```

Figure 13: Modello completo

```
(Dispersion parameter for inverse.gaussian family taken to be 0.01276082)

    Null deviance: 496.19  on 4375  degrees of freedom
Residual deviance:  58.38  on 4343  degrees of freedom
AIC: 17819
```

Figure 14: Modello ridotto

La discriminante utilizzata per la selezione del modello finale viene fornita dalla bontà di adattamentemento di ciascun modello. Evidenza particolare viene fornita dal valore della devianza residuale per i due modelli, in particolare i valori risultano approssimativamente uguali, per cui l'apporto informativo alla log-verosimiglianza delle 11 variabili escluse risulta approssimativamente nullo.
La devianza residuale per entrambi i modelli risulta essere abbastanza bassa, ciò vuol dire che la differenza della log-verosimiglianza del modello saturo ed il valore della log-verosimiglianza del modello stimato è piccola.
Il modello ridotto presenta un valore del criterio di informazione Akaike leggermente inferiore rispetto al modello completo. In generale, la bontà di adattamento dei due modelli risulta essere complessivamente simile, dunque la selezione del modello avviene confrontando i due modelli attraverso il Likelihood Ratio Test.

Abbiamo quindi confrontato tramite il Likelihood Ratio Test (LRT) il modello completo, quello ridotto e quello con il solo termine dell'intercetta.

```
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
     4343     58.380
     4332     58.273 11  0.10733   0.6762
```

Figure 15: Test LRT, confronto modello ridotto e completo

```
Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
     4375     496.19
     4343      58.38 32   437.81 < 2.2e-16 ***
```

Figure 16: Test LRT, confronto modello ridotto e modello intercetta

Secondo il Likelihood Ratio Test, non rifiutiamo l'ipotesi nulla di equivalenza tra il modello completo e quello ridotto (riferimento Figura 15), mentre rifiutiamo l'equivalenza tra il modello che include solo l'intercetta e il ridotto (riferimento Figura 16). Dunque viene la scelta finale del modello ricade su quello ridotto per il principio di parsimonia.

Dopo aver selezionato il modello per entrambe le famiglie di regressioni considerate, è importante comprendere l'influenza dei regressori sul fenomeno che si intende modellare, ovvero il Prezzo.

Di seguito vengono mostrati alcuni dei coefficienti associati alle variabili presenti tra i regressori dei modelli considerati: Le Figure 17 e 18 rappresentano una sintesi dei coefficienti più influenti, sia

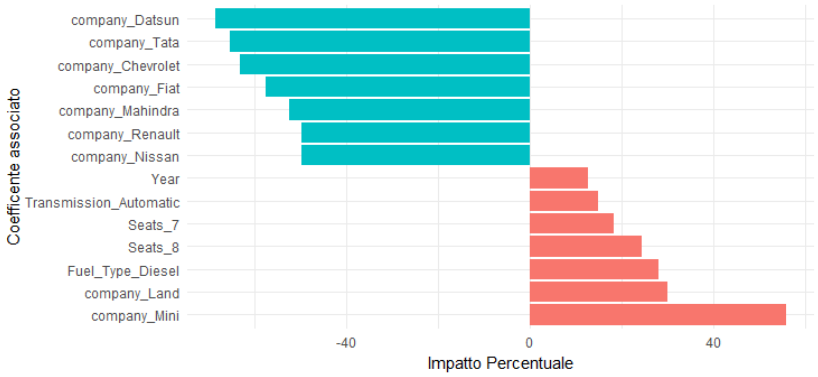

Figure 17: Regressione Gamma

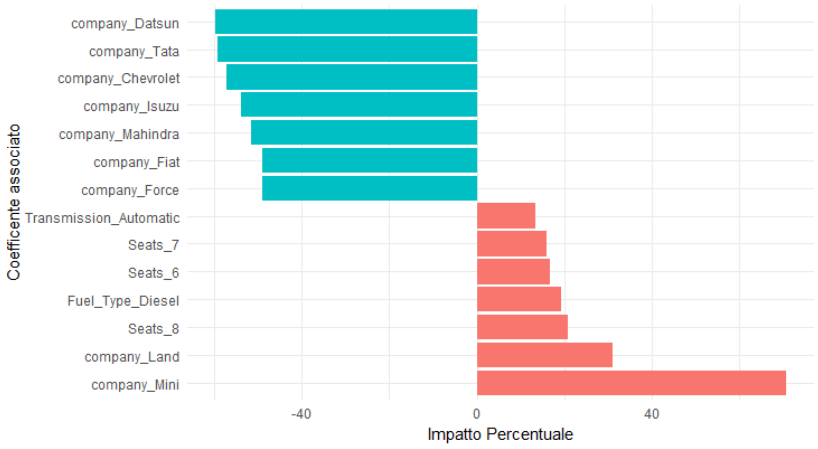

Figure 18: Regressione Gaussiana Inversa

Variabili associate ai coefficienti stimati più influenti sul Prezzo

negativi che positivi per entrambe le regressioni considerate, in termini percentuali, ovvero abbiamo considerato la trasformazione $(exp(\beta_j) - 1) * 100$, che è interpretabile come la variazione percentuale stimata del Prezzo dell'automobile quando la variabile X_j viene aumentata di un'unità, se continua, o quando passa da 0 a 1 se binaria (ceteris paribus). Osservando le due figure si nota come la casa produttrice dell'autovettura abbia un grande impatto sul prezzo dell'automobile usata in entrambe le regressioni, difatti 9 dei 14 coefficenti mostrati, in entrambe le figure, sono associati alla variabile della casa di costruzione.
Un ulteriore dato di interesse riguarda l'influenza del tipo di alimentazione del motore sul prezzo dell'automobile. In particolare, si osserva un incremento di prezzo del 20% (Reg. Gaussiana Inversa) e del 25% (Reg. Gamma) rispetto ad un'automobile alimentata a benzina.
Un altro elemento dell'automobile che pare avere un impatto positivo sul prezzo è il tipo di trasmissione. Nello specifico, l'adozione della trasmissione automatica si traduce in un aumento del prezzo di circa il 17% in entrambi i modelli.
In generale, si evidenzia come i coefficienti più influenti e il relativo impatto risultino sostanzialmente simili per entrambe le regressioni.

In un contesto dove non si conoscono a pieno le dinamiche del fenomeno, può essere utile valutare e decretare il modello più adatto al problema considerato attraverso le perfomance registrate su particolari criteri di informazione.

In particolare per il fenomeno in questione vengono considerati i criteri di informazione Akaike (AIC) e Bayesian (BIC). Basando la scelta del modello più adatto, per descrivere il fenomeno di

Model	**AIC**	**BIC**
Gamma -Rid	16433	16656
Gaussiana Inv. -Rid	17819	18036

Table 1: Criteri di informazione dei modelli

interesse, sui criteri di informazione AIC e BIC, questa ricade sul modello lineare generalizzato nel quale viene ipotizzata una distribuzione Gamma per la componente casuale.

Capacità predittive

Per completezza viene quindi valutata la capacità previsiva dei due modelli considerati sin ora. In un momento antecedente alla fase di stima il set di dati disponibili è stato suddiviso per valutare il comportamento dei due modelli su dati non osservati, in particolare la divisione del dataset è avvenuta come segue:

- il 75% delle osservazioni sono state destinate al *Train*
- il 25% delle osservaizone sono state destinate al *Test*

Così facendo, ottenuta la stima dei modelli sul train, la relativa capacità previsiva dei modelli viene valutata sul test.

Modelli	**MAE**	**MSE**	**RMSE**
Gamma -Rid	2.05	33.90	5.82
Gaussiana Inv. -Rid	2.65	123.83	11.12

Table 2: Performance previsive

Secondo le metriche, risulta evidente come il modello Gamma nel contesto previsivo riesca meglio a cogliere e prevedere le dinamiche del prezzo, difatti questa rispetto alla regressione Gaussiana Inversa minimizza ogni metrica di previsione considerata.

Data le migliori capacità previsive, andiamo quindi a calcolare gli effetti marginali sfruttando il modello Gamma stimato.

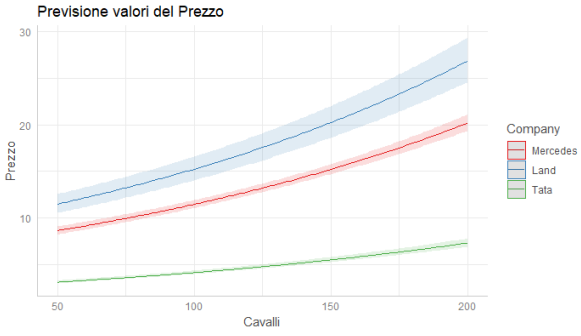

Figure 19: Effetti marignali dei coefficienti delle Case Produttrici Tata,Mercedes,Land e variabile Cavalli

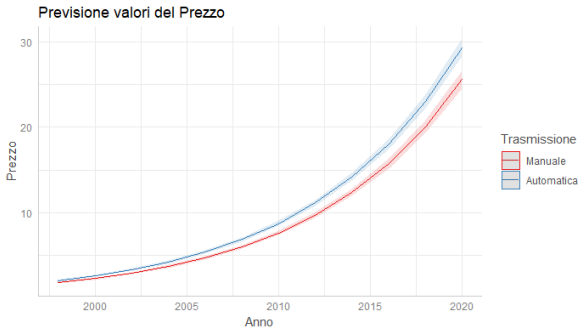

Figure 20: Effetti marginali dei coefficenti della trasmissione e Anno di produzione

Per una questione di sintesi vengono mostrati in figura solo alcuni dei possibili effetti marginali calcolati con i coefficienti del modello Gamma. In particolare seguendo le indicazioni fornite dalla figura 17, sono stati calcolati gli effetti marginali che hanno sul prezzo il numero di cavalli dell'autovettura per le compagnie "Mercedes", "Tata, "Land", fare riferimento alla Figura 19, osserviamo il prezzo

incrementare in modo lineare e simile per le compagnie "Mercedes" e "Land-Rover" all'aumentare del numero di cavalli. Diversamente il prezzo cresce ad un tasso nettamente più lento per la compagnia "Tata" con l'incremento del numero di cavalli.
Ulteriori effetti marginali sono stati calcolati per la variabile binaria "Trasmissione" e la variabile numerica "Anno di produzione", fare riferimento alla Figura 20. In particolare, si osserva un significativo aumento del prezzo in corrispondenza degli anni più recenti, soprattutto per le automobili dotate di trasmissione automatica. Per gli anni di costruzione più remoti, la differenza di prezzo tra le vetture con cambio automatico e manuale risulta essere quasi trascurabile. Continuando con l'analisi degli effetti marginali, nella Figura 21, viene evidenziato l'impatto dei cavalli e del chilometraggio sul prezzo delle autovetture. In particolare, si osserva che per le autovetture con un minor numero di cavalli, il chilometraggio ha un effetto notevolmente ridotto sul prezzo. Al contrario, per le autovetture con un numero di cavalli maggiore, il chilometraggio mostra un impatto negativo più significativo sul valore dell'automobile.

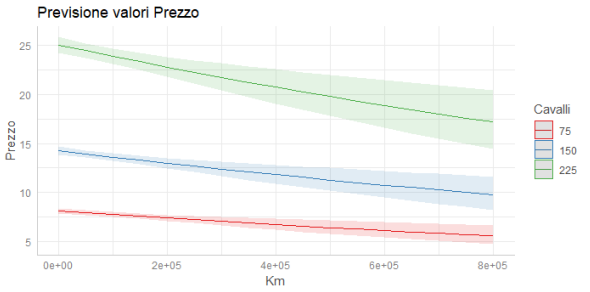

Figure 21

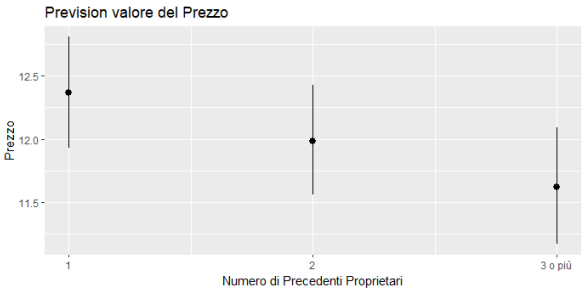

Figure 22

L'ultima considerazione riguarda il numero dei precedenti proprietari e l'incidenza che questa variabile ha sul prezzo dell'automobile, dalla figura 22, notiamo come al crescere del numero di passati proprietari decresca leggermente il prezzo dell'auto.

4 Conclusione

In questo progetto, sono state approfondite le peculiarità e i principali fattori nella determinazione del prezzo mediante l'utilizzo di strumenti di analisi esplorativa e la modellizzazione del fenomeno attraverso due modelli appartenenti alla famiglia dei GLM. Attraverso tali strumenti, abbiamo identificato le sfumature e i fattori che influenzano il prezzo dell'automobile, individuando tra i principali: la casa produttrice, l'anno di produzione, la tipologia di alimentazione del motore e la trasmissione.