Predizione Prezzi Automobili & Classificazione Della Cardiopatia

Julian Prego, Luigi Petraccioli

Contenuti

Abstract

I modelli lineari generalizzati (GLM) rappresentano un'estensione dei modelli di regressione lineare ordinaria, permettendo l'uso di diverse distribuzioni per la variabile di risposta con funzioni di collegamento distintive. Questo progetto esplora le applicazioni pratiche dei GLM utilizzando due set di dati distinti: la "Predizione dei Prezzi delle Auto" e la "Predizione della Cardiopatia". Nel primo set di dati, analizziamo i prezzi delle auto per determinare le variabili che lo influenzano maggiormente. L'obiettivo è creare un modello robusto che possa identificare le caratteristiche fondamentali, come il tipo di carburante, la dimensione del motore e il marchio dell'auto. Questo ci permette di comprendere meglio il mercato automobilistico e di fornire preziose informazioni ai produttori e ai consumatori. Il secondo set di dati riguarda la predizione della cardiopatia, dove il nostro obiettivo è identificare i fattori chiave che contribuiscono al rischio di sviluppare questa condizione. Utilizzando variabili come l'età, il colesterolo e il tipo di dolore toracico, costruiamo un modello che non solo predice il rischio di insufficienza cardiaca, ma evidenzia anche le variabili più critiche per la prevenzione e il trattamento. In entrambe le analisi, ci concentriamo sull'adattamento del modello e sull'identificazione delle variabili più significative piuttosto che sulla pura accuratezza predittiva. Questo approccio permette di ottenere una comprensione più profonda dei fenomeni studiati, fornendo intuizioni che possono guidare decisioni pratiche. Tutte le analisi sono state eseguite utilizzando il linguaggio di programmazione R, sfruttando le sue librerie per la manipolazione dei dati e la modellazione statistica. Questo progetto dimostra come i GLM possano essere applicati a problemi reali, offrendo strumenti analitici avanzati per la comprensione e la risoluzione di problemi complessi in vari campi.

1.Predizione Prezzi Auto

1.1 Introduzione

Uno dei mercati automobilistici più grandi al mondo è quello degli Stati Uniti. Dal 1982, quando Honda ha investito nel mercato automobilistico statunitense, molte altre aziende si sono unite e competono nel mercato automobilistico degli Stati Uniti, risultando in un investimento straniero di oltre 110 miliardi di dollari. Oggi, con lavoratori qualificati, supporti locali e governativi, un enorme mercato dei consumatori e molto altro, il mercato automobilistico degli Stati Uniti è un mercato primario nell'industria automobilistica. Di seguito, il nostro obiettivo è identificare le variabili significative che influenzano il prezzo delle auto e quantificarne l'importanza. Queste analisi di solito vengono eseguite da una terza parte, come una società di consulenza o la divisione di strategia aziendale dell'azienda investitrice. Secondo i nostri risultati, possono manipolare molte variabili, come il design dell'auto, per avere una migliore strategia aziendale per entrare nel mercato automobilistico degli Stati Uniti. Queste analisi possono influenzare direttamente il successo di un investimento di miliardi di dollari. Di conseguenza, le nostre analisi sono cruciali e devono essere dettagliate e valide. Abbiamo scoperto che la distribuzione del prezzo dell'auto è molto simile alla distribuzione Gamma; pertanto, abbiamo utilizzato il GLM con distribuzione Gamma e funzione di collegamento logaritmica per modellare il prezzo delle auto in base a variabili distintive. Abbiamo anche sospettato che fosse possibile modellare il logaritmo del prezzo con distribuzione Gaussiana e funzione di collegamento identità. Tuttavia, la distribuzione del prezzo logaritmico non è vicina alla distribuzione Normale. Di conseguenza, abbiamo utilizzato solo la distribuzione Gamma con la funzione di collegamento logaritmica. Abbiamo eseguito la selezione delle variabili e selezionato il modello più ragionevole. Grazie a queste analisi, siamo stati in grado di identificare diverse variabili significative che contribuiscono al prezzo dell'auto, come il produttore dell'auto, la posizione del motore (le auto con motori posteriori sono solitamente auto sportive con prezzi più alti) e la dimensione del motore (più grande è il motore, più alto è il prezzo). Il nostro set di dati, e di conseguenza le nostre analisi, presentano anche alcune limitazioni. Ad esempio, le auto elettriche

costituiscono più del 8,5% del mercato automobilistico statunitense ma non sono incluse nel nostro set di dati. Inoltre, mancano marchi di lusso come Rolls-Royce e Lincoln. Infine, la maggior parte delle auto sportive è assente nel nostro set di dati, evidenziando la nostra limitazione nell'analizzare le variabili di prezzo delle auto sportive e di lusso. Di seguito, presentiamo una descrizione dettagliata della nostra analisi, dei metodi e dei risultati.

1.2 Caratteristiche Dei Dati

I dati sono stati raccolti dal sito web Kaggle (https://www.kaggle.com/datasets/hellbuoy/car-price-prediction), una comunità online open-source di scienziati dei dati e praticanti di machine learning. I dati non presentavano valori mancanti ed erano pronti per l'analisi. Tuttavia, abbiamo apportato piccole modifiche e correzioni nel dataset. Abbiamo rimosso la colonna CAR ID in quanto non contiene informazioni utili per le nostre analisi. Inoltre, abbiamo cambiato i nomi delle auto nei nomi dei produttori. In questo modo, la variabile rappresenta la reputazione del marchio dell'auto (o del produttore), che potrebbe influenzare il prezzo dell'auto, anziché il modello dell'auto, che è unico per la maggior parte delle auto e non influisce sul prezzo dell'auto. In aggiunta, abbiamo rimosso alcune delle covariate che presentavano problemi di collinearità. Ad esempio, le covariate tipo di motore e numero di cilindri sono correlate, dal momento che i tipi di motore sono solitamente determinati dal numero di cilindri e dal modo in cui questi cilindri sono disposti. Lo stesso scenario si applica ai sistemi di alimentazione e al tipo di carburante, poiché il sistema di alimentazione di un'auto può essere influenzato dal tipo di carburante che consuma. Per risolvere questo problema di collinearità nelle nostre analisi, abbiamo rimosso il tipo di motore e il sistema di alimentazione. Successivamente, abbiamo cercato di determinare la distribuzione della risposta da utilizzare per la funzione di collegamento e la famiglia di distribuzione adeguate. La distribuzione della risposta assomiglia alla distribuzione Gamma (Figura 1 (A)). Abbiamo anche visualizzato il logaritmo della risposta, poiché potrebbe essere gaussiano. Come si può vedere dalla Figura 1 (B), il logaritmo del prezzo non assomiglia alla distribuzione gaussiana. Pertanto, abbiamo deciso

di utilizzare solo la distribuzione Gamma con la funzione di collegamento logaritmico (Vedi ulteriori dettagli nelle prossime sezioni).

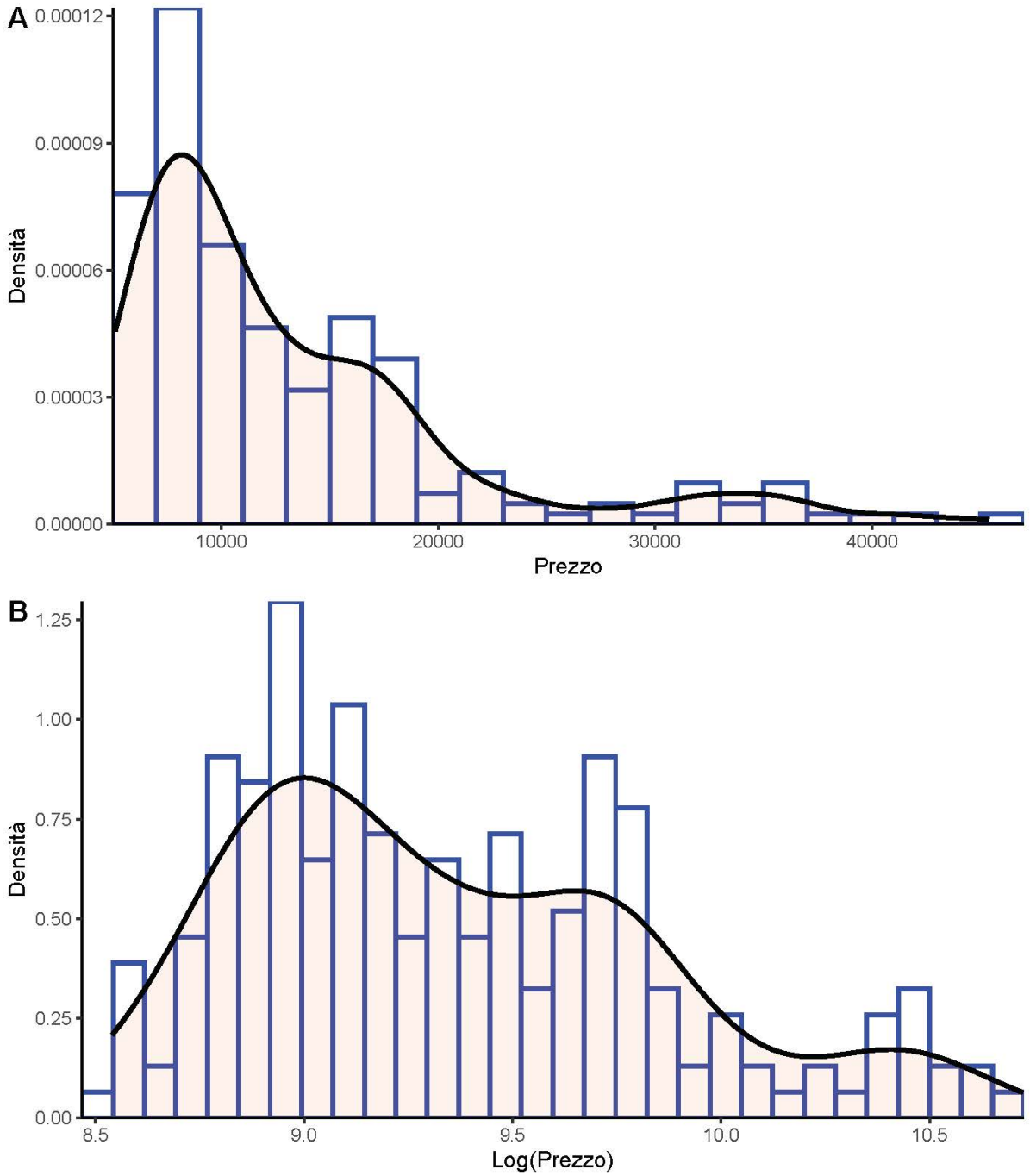

Figura 1: Distribuzione dei prezzi. A)Risposte B)Logaritmo delle risposte.

La Figura 2 mostra una panoramica dei produttori, dell'intervallo dei prezzi delle loro auto e del tipo di carburante delle loro produzioni. Come è evidente,

mancano le auto elettriche e le auto diesel sono una minoranza. Inoltre, come mostrato nella Figura 2, i marchi delle auto (o i produttori) possono influenzare il prezzo dell'auto. Ad esempio, le auto Porsche hanno un prezzo più alto rispetto alle auto Nissan o Mazda. Inoltre, si nota l'assenza di famosi marchi di auto sportive come Ferrari e produttori di lusso come Rolls-Royce e Lincoln.

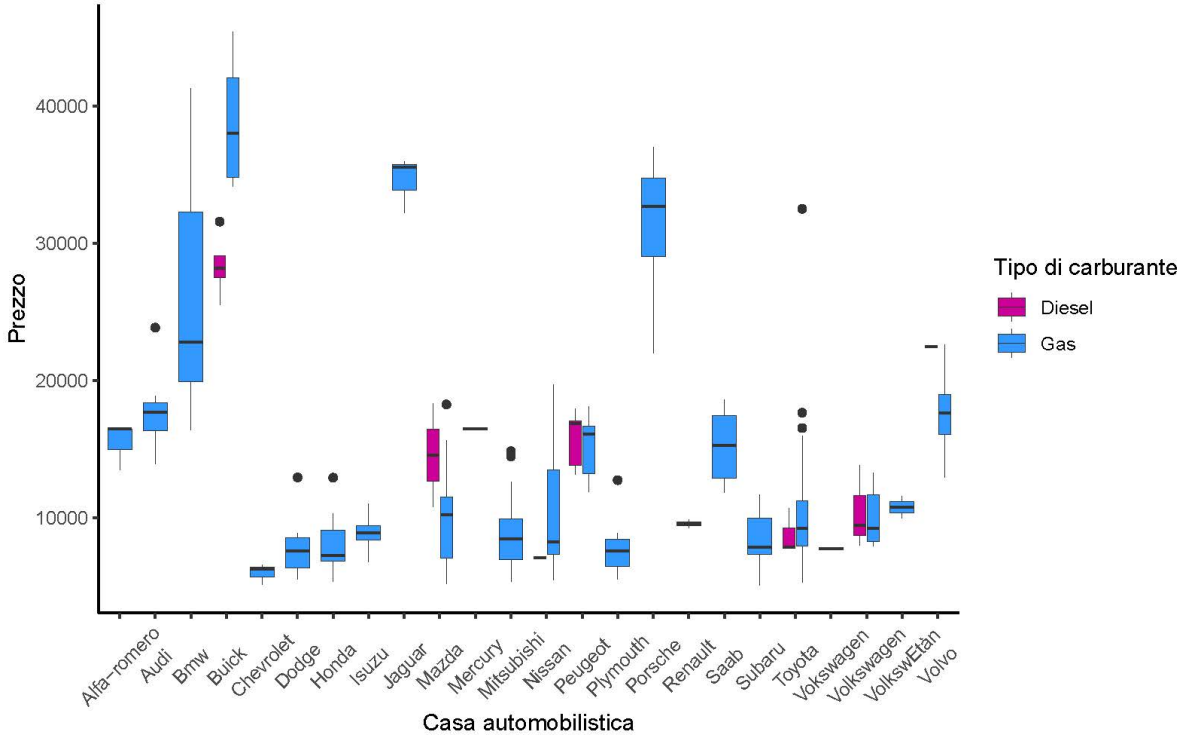

Figura 2: Range dei prezzi delle auto nei differenti brande tipi di carburante.

Nella Figura 3, si può vedere che il prezzo dell'auto è correlato alla dimensione del motore; tuttavia, potrebbe non essere una correlazione lineare. Inoltre, ciò che risalta in questa figura è la crescita generale della dimensione del motore con l'aumento del numero di cilindri, e anche la risposta aumenterà con l'aumento di uno qualsiasi di essi.

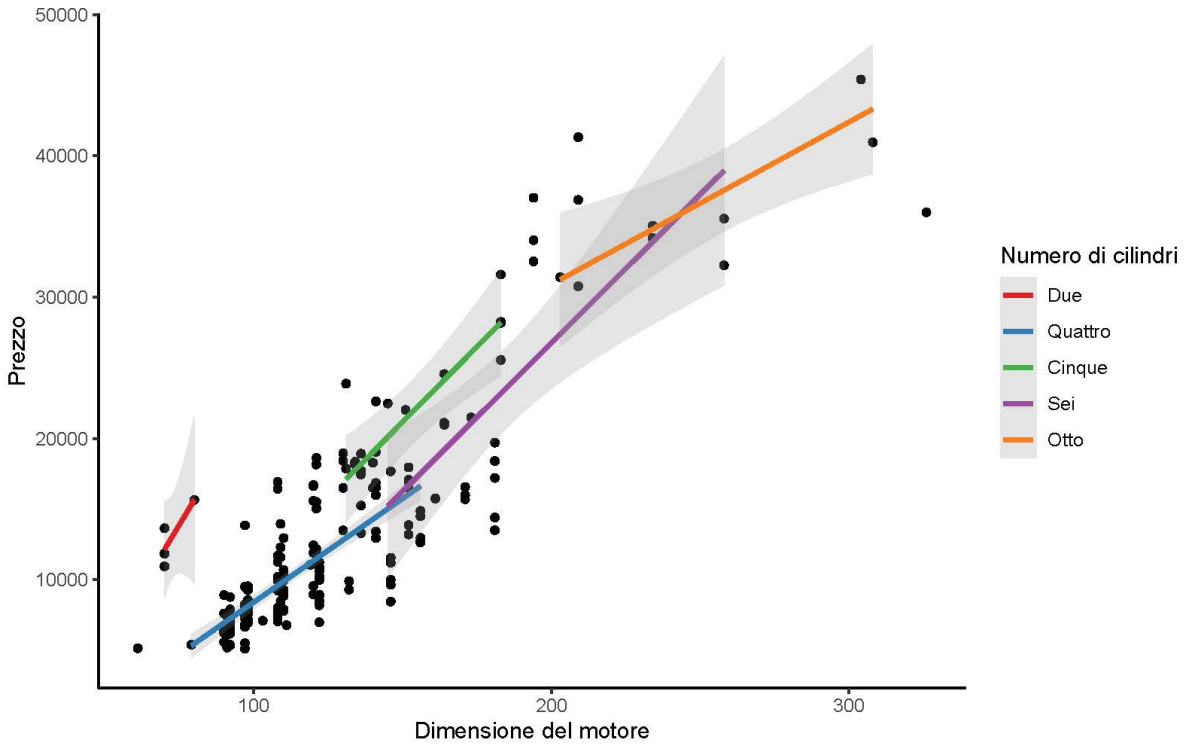

Figura 3: Correlazione fra taglia del motore e prezzo per diversi numeri di cilindri.

Abbiamo anche sospettato che potrebbero esserci delle tendenze con l'incremento quadratico delle variabili numeriche. Pertanto, abbiamo indagato su questi pattern. Ad esempio, nella Figura 4, abbiamo diviso il passo ruota delle auto in quattro gruppi e visualizzato la tendenza tra il passo ruota e la risposta. Come mostra questa figura, il prezzo dell'auto non sta crescendo linearmente con l'aumento del passo ruota. Pertanto, abbiamo incluso anche termini quadratici nel nostro modello statistico. Nella sezione successiva, vengono descritti i dettagli delle nostre analisi statistiche.

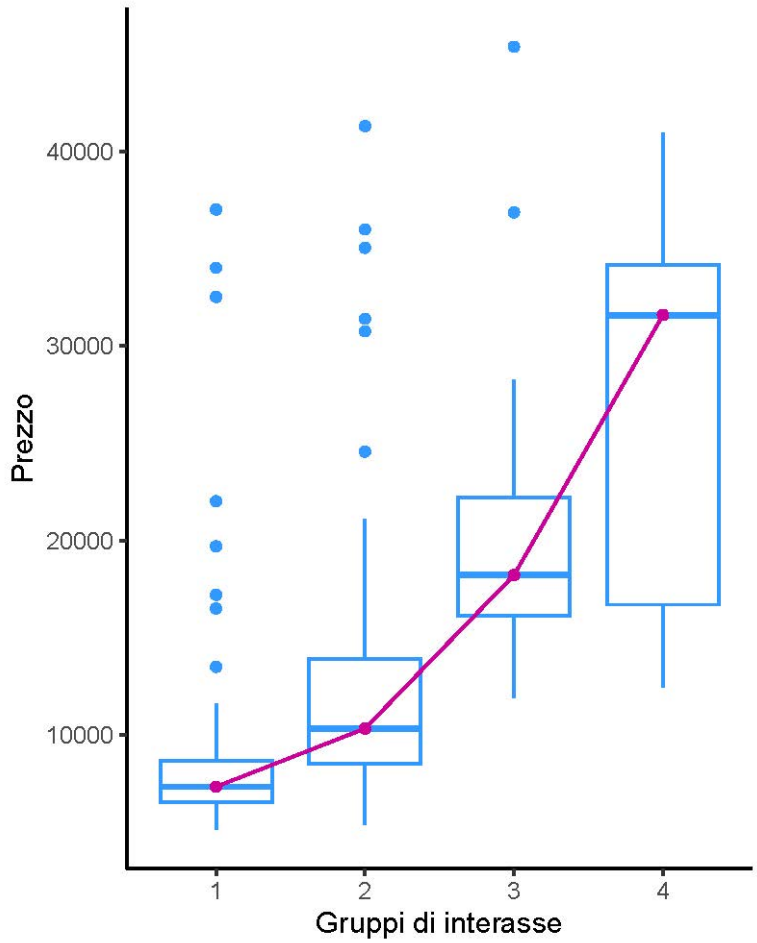

Figura 4: L'aumento quadratico del prezzo con diversi gruppi di interasse.

CODICE

```
library(readr)

car_dat <- read_csv("C:/Users/Juli/Desktop/Progetto/data/CarPrice_Assignment.csv")

test_string<-toString(car_dat$CarName[1])

strsplit(test_string,"-")

test_string2<-toString(car_dat$CarName[4])

test_vec<-c(strsplit(test_string2," "))

test_car_comp<-c(rep(NA,nrow(car_dat)))

test_car_list<-vector(mode = "list", length = nrow(car_dat))

for(i in 1:nrow(car_dat)){
```

```
  test_string<-toString(car_dat$CarName[i])
  test_car_list[[i]]<-c(strsplit(test_string," "))
  test_car_pre<-unlist(test_car_list[[i]][1])
  test_car_comp[i]<-test_car_pre[1]}
# Crea un nuovo vettore per i nomi delle compagnie corrette.
car_comp_final<-c(rep(NA,nrow(car_dat)))
# Ciclo for per correggere i nomi delle compagnie con errori di battitura.
for(i in 1:nrow(car_dat)){
  if(test_car_comp[i]=="maxda"){
    car_comp_final[i]<-"mazda"  }
  else if(test_car_comp[i]=="Nissan"){
    car_comp_final[i]<-"nissan"  }
  else if(test_car_comp[i]=="porcshce"){
    car_comp_final[i]<-"toyota"  }
  else if(test_car_comp[i]=="vokswagen"){
    car_comp_final[i]<-"volkswagen"}
  else if(test_car_comp[i]=="vw"){
    car_comp_final[i]<-"volkswagen"}
  else if(test_car_comp[i]=="toyouta"){
    car_comp_final[i]<-"toyota" }
  else{
    car_comp_final[i]<-test_car_comp[i]  }}
# Funzione per mettere la prima lettera in maiuscolo.
firstup <- function(x) {
  substr(x, 1, 1) <- toupper(substr(x, 1, 1))  x}
# Applica la funzione firstup a car_comp_final per mettere in maiuscolo la prima
lettera di ogni nome di compagnia.
```

```
car_comp_final<-firstup(car_comp_final)

# Aggiunge la colonna car_company al dataset car_dat con i nomi delle compagnie corretti.

car_dat$car_company<-car_comp_final

# Crea un istogramma del prezzo delle macchine.

hist(car_dat$price)

# possiamo usare la regressione esponenziale o il modello log-lineare

# rimuovi car id e car name poiché non sono utili nell'analisi

# car name si sovrappone alle variabili car company

car_dat2<-car_dat[,-c(1,3)]

#estrai variabili di tipo carattere

character_var<-car_dat2[, sapply(car_dat2, class) == 'character']

character_var[sapply(character_var, is.character)] <- lapply(character_var[sapply(character_var, is.character)], as.factor)

#estrai variabili non di tipo carattere

no_chara<-car_dat2[, sapply(car_dat2, class) != 'character']

#summary(no_chara)

car_dat3<-cbind(character_var,no_chara)

## grafici di distribuzione

p1 <- ggplot(car_dat3) + aes(x =price) +

 geom_histogram(aes(y=..density..), size=1,color="blue", fill="white", binwidth = 2000)+

 geom_density(alpha=.05, fill="red", size = 1) +

 scale_x_continuous(expand = c(0, 0)) +

 scale_y_continuous(expand = c(0, 0)) +

 theme_classic() +

 theme(line = element_line(size = 0.5)) +
```

```
   xlab('Prezzo') + ylab('Densità')
p2 <- ggplot(car_dat3) + aes(x = log(price)) +
 geom_histogram(aes(y=..density..), size=1,color="blue", fill="white")+
 geom_density(alpha=.05, fill="red", size = 1) +
 scale_x_continuous(expand = c(0, 0)) +
 scale_y_continuous(expand = c(0, 0)) +
 theme_classic() +
 theme(line = element_line(size = 0.5))+
   xlab('Log(Prezzo)') + ylab('Densità')
p <-
 ggpubr::ggarrange(p1, p2,
                   labels = c("A", "B"),
                   ncol = 1, nrow = 2)
p
ggsave('C:/Users/Juli/Desktop/Progetto/Figures/distribuzione_prezzo_auto.pdf
', p, width = 7, height = 8)
## Box plots
p <-
 ggplot(car_dat3, aes(x=car_company, y=price, fill=fueltype)) +
 geom_boxplot(lwd=0.25) +
 xlab('Casa automobilistica') + ylab('Prezzo')+
 scale_fill_manual(values=c("#cc0099", "#3399ff"),
                    name = "Tipo di carburante",
                    labels = c("Diesel", "Gas"))+
 theme_classic() +
 theme(axis.text.x = element_text(angle = 45,vjust = 0.5))
```

```
ggsave('C:/Users/Juli/Desktop/Progetto/Figures/distribuzione_casa_automobilistica.pdf', p, width = 8, height = 5)

car_dat3$cylindernumber <- factor(car_dat3$cylindernumber, levels=c('two', 'three', 'four', 'five', 'six', 'eight', 'twelve'),
                                  labels=c('Due', 'Tre', 'Quattro', 'Cinque', 'Sei', 'Otto', 'Dodici'))

p <-
  ggplot(car_dat3) +
  aes(x = enginesize, y = price, color = cylindernumber) +
  geom_point(color = "black") +
  geom_smooth(method = "lm", alpha = 0.2) +
  xlab('Dimensione del motore') + ylab('Prezzo')+
  scale_colour_brewer(palette = "Set1", name = "Numero di cilindri")+
  theme_classic()

ggsave('C:/Users/Juli/Desktop/Progetto/Figures/dimensione_motore.pdf', p, width = 8, height = 5)

res <- car_dat3 %>% mutate(category=cut(wheelbase, breaks=4, labels=c("1","2","3", "4")))

p <-
 ggplot(res) +
 aes(x = category, y = price) +
 geom_boxplot(color = "#3399ff") + stat_summary(fun=median, geom="line", aes(group=1), color = "#cc0099") +
 stat_summary(fun=median, geom="point", color = "#cc0099")+ #ggtitle('Leffetto quadratico dell'interasse sul prezzo')+
 xlab('Gruppi di interasse') + ylab('Prezzo')+
 theme_classic()

p
```

ggsave('C:/Users/Juli/Desktop/Progetto/Figures/gruppi_interasse.pdf', p, width = 4, height = 5)

1.3 Analisi Statistica

Nel dataset dei prezzi delle auto, abbiamo 205 osservazioni e 27 variabili, inclusa la variabile di risposta (prezzo). Prima di iniziare l'analisi statistica, definiamo il dizionario dei dati per comprendere meglio il nostro dataset.

Variable Name	Definition
CarID	ID univoco di ciascuna osservazione (intero)
Symboling	La valutazione del rischio assicurativo assegnata, un valore di +3 indica che la macchina è rischiosa, -3 che probabilmente è abbastanza sicura. (Categorica)
car name	Nome della macchina (Categorica)
car company	Nome della casa automobilistica (Categorica)
fueltype	Tipo di carburante per macchina, ad esempio gas o diesel (categoriale)
aspiration	Aspirazione utilizzata in una macchina (Categorica)
doornumber	Numero di porte in una macchina(categoriale)
carbody	carrozzeria della macchina (Categorica)
drivewheel	tipo di ruota motrice (Categorica)
enginelocation	Posizione del motore della macchina (Categorico)
wheelbase	Passo della macchina (numerico)
carlength	Lunghezza della macchina (numerico)
carwidth	Larghezza della macchina(numerico)
carheight	altezza della macchina (numerico)
curbweight	Il peso di un'auto senza occupanti né bagagli. (Numerico)
enginetype	Tipo di motore. (Categorico)
cylindernumber	cilindro posizionato nella macchina (Categorico)
enginesize	Dimensioni della macchina (numerico)
fuelsystem	Sistema di alimentazione della macchina (Categorico)
boreratio	Borerazione della macchina (numerico)
stroke	Corsa o volume all'interno del motore (Numerico)
compressionratio	rapporto di compressione della macchina (numerico)

horsepower	Potenza (numerica)
peakrpm	RPM di picco della macchina (numerico)
citympg	Chilometraggio in città (numerico)
highwaympg	Chilometraggio in autostrada (numerico)
price	Prezzo della macchina (numerico)

Tabella 1: Dizionario dei dati.

Decidiamo innanzitutto quale modello utilizzare per l'analisi esaminando gli istogrammi del prezzo delle auto e del log(prezzo delle auto). Non possiamo utilizzare la regressione lineare normale per modellare il prezzo delle auto poiché né il prezzo delle auto né il log(prezzo delle auto) seguono una distribuzione normale. Dopo un'ulteriore ispezione, abbiamo deciso di utilizzare la regressione gamma con log-link poiché sia il prezzo delle auto che il log(prezzo delle auto) seguono una distribuzione gamma. Abbiamo specificamente scelto il log-link per la regressione gamma invece del suo link canonico (link negativo-inverso) poiché il prezzo delle auto può essere solo positivo.

Per l'analisi del dataset dei prezzi delle auto, abbiamo deciso di costruire due modelli:

1) Il modello ad effetto principale
2) Il modello ad effetto principale più la covarianza al quadrato

Abbiamo deciso di non includere termini di interazione nella nostra analisi poiché la maggior parte dei nostri covariati sono categorici e la maggior parte di questi covariati categorici ha più di 2 livelli.

Per il primo modello, abbiamo costruito il modello con tutti i covariati (escludendo i covariati che causano problemi di collinearità) con il prezzo come risposta. La formula per il modello è la seguente.

$$\log(\mu) = \beta_0 + \sum_{i=1}^{22} \beta_0 x_1$$

La formula per il modello completo ad effetto principale è:

Formula del Primo Modello

log(prezzo) ~ tipo di carburante + aspirazione + numero di porte + tipo di carrozzeria + tipo di trazione + posizione del motore + numero di cilindri + casa automobilistica + simbolizzazione + passo + lunghezza dell'auto + larghezza

dell'auto + altezza dell'auto + peso a vuoto + dimensione del motore + rapporto alesaggio-corsa + corsa + rapporto di compressione + potenza del motore + giri al minuto massimi + consumo in città + consumo in autostrada

La seguente tabella mostra i risultati del modello completo ad effetto principale.

	Stime	Pr(>|t|)	Significatività
(Intercept)	7.5	8.47e-12	***
Fueltype gas	-0.23	0.54	not significant
Aspiration Turbo	0.08	0.10	not significant
Doornumber two	-0.03	0.28	not significant
Carbody Hardtop	-0.16	0.05	*
Carbody Hatchback	-0.21	1.87e-03	**
Carbody Sedan	-0.14	0.06	.
Carbody Wagon	-0.15	0.06	.
Drivewheel fwd	-0.04	0.52	not significant
Drivewheel rwd	-6.23e-03	0.93	not significant
enginelocation rear	0.78	6.63e-09	***
cylinder num5	0.04	0.75	not significant
cylinder num4	0.20	0.19	not significant
cylinder num6	0.04	0.75	not significant
cylinder num3	0.51	0.02	*
cylinder num12	-0.13	0.66	not significant
cylinder num2	0.31	0.12	not significant
Audi	0.07	0.62	not significant
BMW	0.36	8.25e-04	***
Buick	-0.05	0.72	not significant
Chevrolet	-0.25	0.05	.
Dodge	-0.35	1.07e-03	**
Honda	-0.18	0.09	.
Isuzu	-0.16	0.30	not significant
Jaguar	-0.37	0.02	*
Mazda	-0.09	0.34	not significant
Mercury	-0.15	0.32	not significant
Mitsubishi	-0.40	2.21e-04	***
Nissan	-0.15	0.12	not significant
Peugeot	-0.37	4.53e-03	**
Plymouth	-0.36	8.23e-04	***
Porsche	0.03	0.83	not significant
Renault	-0.27	0.05	.
Saab	0.1	0.39	not significant

Subaru	-0.20	0.10	.
Toyota	-0.20	0.03	*
Volkswagen	-0.12	0.26	not significant
Volvo	-0.09	0.43	not significant
symboling	1.43e-03	0.93	not significant
wheelbase	0.02	6.54e-04	***
carlength	-7.39e-03	0.02	*
carwidth	0.03	0.02	*
carheight	-0.03	1.06e-04	***
curbweight	5.29e-04	1.21e-06	***
enginesize	2.55e-03	0.07	.
boreratio	-0.17	0.1	.
stroke	-0.01	0.82	not significant
compressionratio	-0.01	0.62	not significant
horsepower	0.02e-04	0.41	not significant
peakrpm	5.31e-05	0.15	not significant
citympg	-0.01	0.08	.
highwaympg	0.01	0.19	not significant

Tabella 2: Risultati completi del modello a effetto principale.
*Legenda: eccessivo *** forte ** moderato * borderline .*

Dalla Tabella 2, possiamo vedere che alcuni covariati sono insignificanti. Pertanto, abbiamo condotto una selezione stepwise in entrambe le direzione utilizzando l'AIC come criterio di selezione. L'AIC, o criterio di informazione di Akaike, è un metodo popolare per la selezione delle variabili per la costruzione del modello.

Per scegliere un modello da una sequenza di modelli candidati $M_i,\ i = 1,\ 2,\ \ldots,\ K$, l'AIC è definito come:

$$AIC_i = -2\ log\ L_i + 2V_i$$

Dalla selezione stepwise con l'AIC, abbiamo ottenuto il modello stepwise per gli effetti principali.

La formula per il modello stepwise ad effetto principale è la seguente:

Formula del Primo Modello Stepwise

log(prezzo) ~ aspirazione + tipo di carrozzeria + posizione del motore + numero di cilindri + casa automobilistica + passo + lunghezza dell'auto + larghezza dell'auto + altezza dell'auto + peso a vuoto + dimensione del motore + rapporto

alesaggio-corsa + giri al minuto massimi + consumo in città + consumo in autostrada

La seguente tabella mostra i risultati della selezione stepwise per gli effetti principali.

	Stime	Pr(>|t|)	Significatività
(Intercept)	7.12	3.04e-14	***
Aspiration Turbo	0.11	2.93e-04	***
Carbody Hardtop	-0.11	0.05	.
Carbody Hatchback	-0.19	2.73e-03	**
Carbody Sedan	-0.10	0.12	not significant
Carbody Wagon	-0.12	0.12	not significant
enginelocation rear	0.82	9.36e-11	***
cylinder num5	0.02	0.84	not significant
cylinder num4	0.17	0.2	not significant
cylinder num6	0.01	0.92	not significant
cylinder num3	0.49	0.02	*
cylinder num12	-0.13	0.46	not significant
cylinder num2	0.19	0.09	.
Audi	0.02	0.84	not significant
BMW	0.34	5.85e-04	***
Buick	-1.1	0.39	not significant
Chevrolet	-0.29	0.02	*
Dodge	-0.4	2.17e-05	***
Honda	-0.23	-0.01	*
Isuzu	-0.15	0.13	not significant
Jaguar	-0.45	1.25e-03	***
Mazda	-0.12	0.20	not significant
Mercury	-0.16	0.30	not significant
Mitsubishi	-0.45	8.02e-07	***
Nissan	-0.18	0.03	*
Peugeot	-0.39	4.68e-04	***
Plymouth	-0.40	2.5e-05	***
Porsche	0.02	0.90	not significant
Renault	-0.32	5.46e-03	**
Saab	0.07	0.55	not signicant
Subaru	-0.12	0.04	*
Toyota	-0.22	9.97e-03	**
Volkswagen	-0.15	0.08	.
Volvo	-0.12	0.25	not significant

wheelbase	0.02	1.83e-04	***
carlength	-0.01	5.63e-03	**
carwidth	0.03	0.02	*
carheight	-0.03	1.43e-05	***
curbweight	6.01e-4	5.71e-12	***
enginesize	2.76e-3	0.02	*
boreratio	-0.16	0.09	.
peakrpm	6.5e-5	0.03	*
citympg	-0.02	0.02	*
highwaympg	0.01	0.09	.

Tabella 3: Risultati del Modello Stepwise degli Effetti Principali.
*Legenda: eccessivo *** forte ** moderato * borderline .*

Per il secondo modello, abbiamo costruito il modello con tutti i covariati (escludendo i covariati che causano problemi di collinearità) più i termini numerici al quadrato.

Formula del Secondo Modello Completo

log(prezzo) ~ tipo di carburante + aspirazione + numero di porte + tipo di carrozzeria + tipo di trazione + posizione del motore + numero di cilindri + casa automobilistica + simbolizzazione + passo + lunghezza dell'auto + larghezza dell'auto + altezza dell'auto + peso a vuoto + dimensione del motore + rapporto alesaggio-corsa + corsa + rapporto di compressione + potenza del motore + giri al minuto massimi + consumo in città + consumo in autostrada + passo^2 + lunghezza dell'auto^2 + larghezza dell'auto^2 + altezza dell'auto^2 + peso a vuoto^2 + dimensione del motore^2 + rapporto alesaggio-corsa^2 + giri al minuto massimi^2 + consumo in città^2 + consumo in autostrada^2

La seguente tabella mostra i risultati del modello completo ad effetto principale con i termini numerici al quadrato.

	Stime	Pr(>|t|)	Significatività
(Intercept)	18.87	0.20	not significant
Fueltype gas	0.21	0.63	not significant
Aspiration Turbo	0.10	0.05	.
Doornumber two	-0.03	0.29	not significant
Carbody Hardtop	-0.19	0.02	*
Carbody Hatchback	-0.22	2.36e-03	***
Carbody Sedan	-0.15	0.05	*
Carbody Wagon	-0.14	0.11	not significant
Drivewheel fwd	-0.02	0.67	not significant

Drivewheel rwd	-0.02	0.74	not significant
enginelocation rear	0.75	4.54e-07	***
cylinder num5	0.12	0.47	not significant
cylinder num4	0.26	0.19	not significant
cylinder num6	0.10	0.47	not significant
cylinder num3	0.56	0.04	*
cylinder num12	-0.41	0.21	not significant
cylinder num2	0.16	0.58	not significant
Audi	0.05	0.78	not significant
BMW	0.42	1.13e-03	**
Buick	0.18	0.38	not significant
Chevrolet	-0.25	0.09	.
Dodge	-0.36	2.85e-03	**
Honda	-0.19	0.11	not significant
Isuzu	-0.10	0.43	not significant
Jaguar	-0.04	0.87	not significant
Mazda	-0.02	0.83	not significant
Mercury	-0.09	0.61	not significant
Mitsubishi	-0.40	7.56e-04	***
Nissan	-0.10	0.37	not significant
Peugeot	-0.22	0.14	not significant
Plymouth	-0.37	1.85e-03	**
Porsche	0.06	0.68	not significant
Renault	-0.25	0.08	.
Saab	0.14	0.30	not significant
Subaru	-0.13	0.35	not significant
Toyota	-0.18	0.10	.
Volkswagen	-0.11	0.36	not significant
Volvo	1.64e-03	0.99	not significant
symboling	2.92e-03	0.87	not significant
wheelbase	0.04	0.65	not significant
carlength	0.02	0.73	not significant
carwidth	-0.25	0.52	not significant
carheight	-0.22	0.37	not significant
curbweight	1.59e-03	6.92e-03	**
enginesize	9.74e-04	0.81	not significant
boreratio	-0.11	0.94	not significant
stroke	0.22	0.74	not significant
compressionratio	0.02	0.63	not significant
horsepower	7.68e-04	0.53	not significant
peakrpm	-5.02e-04	0.21	not significant
citympg	-0.04	0.29	not significant

highwaympg	7.74e-04	0.98	not significant
I(wheelbase^2)	-9.46e-05	0.84	not significant
I(carlength^2)	-7.12e-05	0.59	not significant
I(carwidth^2)	2.13e-03	0.47	not significant
I(carheight^2)	1.67e-03	0.46	not significant
I(curbweight^2)	-2.03e-07	0.06	.
I(enginesize^2)	5.67e-06	0.47	not significant
I(boreratio^2)	-0.02	0.95	not significant
I(peakrpm) ^2	5.46e-08	0.16	not significant
I(citympg^2)	4.46e-04	0.46	not significant
I(highwaympg^2)	1.25e-04	0.81	not significant

Tabella 4: Risultati del Modello Completo degli Effetti Principali con Termini al Quadrato.
*Legenda: eccessivo *** forte ** moderato * borderline.*

Dalla Tabella 4, possiamo vedere che alcuni covariati sono insignificanti. Pertanto, abbiamo condotto una selezione stepwise in entrambe le direzioni utilizzando l'AIC come criterio di selezione. Dalla selezione stepwise con l'AIC, abbiamo ottenuto il modello stepwise per gli effetti principali più i termini numerici al quadrato.

La formula per il modello stepwise ad effetto principale più i termini numerici è la seguente:

Formula del Secondo Modello Stepwise

log(prezzo) ~ aspirazione + tipo di carrozzeria + posizione del motore + casa automobilistica + altezza dell'auto + peso a vuoto + giri al minuto massimi + consumo in città + passo^2 + lunghezza dell'auto^2 + larghezza dell'auto^2 + peso a vuoto^2 + dimensione del motore^2 + giri al minuto massimi^2 + consumo in città^2 + consumo in autostrada^2

La seguente tabella mostra i risultati della selezione stepwise per gli effetti principali più i termini numerici al quadrato.

	Stime	Pr(>|t|)	Significatività
(Intercept)	9.07	6.38e-13	***
Aspiration Turbo	0.11	7.62e-05	***
Carbody Hardtop	-0.17	0.01	*
Carbody Hatchback	-0.23	1.82e-04	***
Carbody Sedan	-0.16	0.01	*

Carbody Wagon	-0.16	0.02	*
enginelocation rear	0.68	5.36e-11	***
Audi	-0.03	0.78	not significant
BMW	0.31	5.16e-04	***
Buick	0.04	0.73	not significant
Chevrolet	-0.19	0.07	.
Dodge	-0.35	1.1e-03	**
Honda	-0.2	0.02	*
Isuzu	-0.13	0.18	not significant
Jaguar	-0.12	0.46	not significant
Mazda	-0.10	0.20	not significant
Mercury	-0.17	0.23	not significant
Mitsubishi	-0.39	3.90e-06	***
Nissan	-0.16	0.05	*
Peugeot	-0.33	7.12e-04	***
Plymouth	-0.35	8.51e-05	***
Porsche	0.02	0.86	not significant
Renault	-0.29	0.01	**
Saab	0.08	0.43	not significant
Subaru	-0.25	2.34e-03	**
Toyota	-0.22	4.98e-03	**
Volkswagen	-0.15	0.07	.
Volvo	-0.1	0.27	not significant
carheight	-0.03	2.61e-05	***
curbweight	1.43e-03	2.19e-05	***
peakrpm	-4.85e-04	0.14	not significant
citympg	-0.04	1.13e-03	**
I(wheelbase^2)	1.06e-04	2.49e-05	***
I(carlength^2)	-2.17e-05	7.87e-03	**
I(carwidth^2)	1.98e-04	0.02	*
I(curbweight^2)	-1.71e-07	5.98e-03	**
I(enginesize^2)	3.96-06	0.01	*
I(peakrpm^2)	5.27e-08	0.09	.
I(citympg^2)	3.5e-04	0.08	.
I(highwaympg^2)	1.84e-04	0.06	.

Tabella 5: Risultati del Modello Stepwise degli Effetti Principali con Termini al Quadrato.
*Legenda: eccessivo *** forte ** moderato * borderline.*

Per ogni modello stepwise, abbiamo deciso di condurre un test del rapporto di verosimiglianza per confrontarlo con i rispettivi modelli completi. Di seguito sono riportati i risultati dei test del rapporto di verosimiglianza.

H_0: $Il\ nostro\ modello\ stepwise$ è $lo\ stesso\ del\ rispettivo\ modello\ completo$.

H_1: $Il\ nostro\ modello\ stepwise$ è $diverso\ dal\ rispettivo\ modello\ completo$.

Tests	Modello	LogVerosimiglianza Valore	Statistica Test	$P(\chi^2 \geq \chi^2_{test})$
Primo Modello Completo vs Primo Modello Stepwise	PMC PMS	-1733 (df=53) -1735 (df=45)	1.93 (df=8)	0.983
Secondo Modello Completo vs Secondo Modello Stepwise	SMC SMS	-1721 (df=63) -1731 (df=41)	9.42 (df=22)	0.991

Tabella 6: Test del Rapporto di Verosimiglianza per Modelli Completi vs Modelli AIC Stepwise.

Poiché i p-valori per entrambi i test del rapporto di verosimiglianza sono elevati, non possiamo rifiutare l'ipotesi nulla. Inoltre, poiché i p-valori sono vicini a 1, ci sono prove significative per supportare che i nostri modelli stepwise sono gli stessi dei rispettivi modelli completi.

Successivamente, effettueremo test di devianza per entrambi i migliori modelli al fine di verificare l'adeguatezza del modello.

$$H_0: Il\ nostro\ modello\ stepwise \text{ è } adeguato.$$

$$H_a: Il\ nostro\ modello\ stepwise\ non \text{ è } adeguato.$$

Modello	Devianza Residua	$P(\chi^2 \geq \chi^2_{test})$
Primo Modello Stepwise	2.01 (161 df)	1
Secondo Modello Stepwise	1.93 (165 df)	1

Tabella 7: Test della Devianza per Modelli AIC Stepwise.

Poiché i p-valori per entrambi i test di devianza sono elevati, non possiamo rifiutare l'ipotesi nulla. Inoltre, poiché i p-valori sono 1, ci sono prove significative per supportare che i nostri modelli stepwise sono adeguati.

Abbiamo quindi deciso di scegliere il miglior modello tra i due modelli stepwise dopo il test di adeguatezza del modello. Poiché il Primo Modello Stepwise e il Secondo Modello Stepwise non sono modelli nidificati l'uno per l'altro, non possiamo utilizzare il Test del Rapporto di Verosimiglianza per il confronto dei modelli. Invece, abbiamo deciso di utilizzare il punteggio AIC, il Pseudo R^2 e i valori di devianza come criteri di selezione.

Primo Modello Stepwise: AIC=3560,Pseudo-R^2= 0.965,devianza=2.01(161 df)

Secondo Modello Stepwise: AIC=3543,Pseudo-R^2=0.966,devianza=1.93(165 df)

Poiché il Secondo Modello Stepwise ha l'AIC più basso, il Pseudo R^2 più alto e un punteggio di devianza più basso rispetto al Primo Modello Stepwise, abbiamo deciso di scegliere il Secondo Modello Stepwise come il nostro miglior modello.

Successivamente, abbiamo costruito l'intervallo di t-confidenza al 95% per le nostre stime dei parametri. Abbiamo ottenuto l'intervallo di confidenza delle stime dei parametri da

$$\hat{\beta}_i \pm SE_{\hat{\beta}_i} t_{\alpha/2,165}$$

	Stime	SE	Limite inferiore	Limite Superiore
(Intercept)	9.07	1.16	6.77	11.36
Aspiration Turbo	0.11	0.03	0.06	0.17
Carbody Hardtop	-0.17	0.07	-0.31	-0.04
Carbody Hatchback	-0.23	0.06	-0.34	-0.11
Carbody Sedan	-0.16	0.06	-0.28	-0.03
Carbody Wagon	-0.16	0.07	-0.29	-0.02
enginelocation rear	0.68	0.1	0.49	0.87
Audi	-0.03	0.09	-0.21	0.16
BMW	0.31	0.09	0.14	0.48
Buick	0.04	0.12	-0.19	0.27
Chevrolet	-0.19	0.11	-0.40	0.02
Dodge	-0.35	0.09	-0.52	-0.17
Honda	-0.2	0.09	-0.37	-0.03
Isuzu	-0.13	0.09	-0.31	0.06
Jaguar	-0.12	0.16	-0.42	0.19
Mazda	-0.10	0.08	-0.26	0.05

Mercury	-0.17	0.13	-0.44	0.10
Mitsubishi	-0.39	0.08	-0.55	-0.23
Nissan	-0.16	0.08	-0.32	-2.64e-03
Peugeot	-0.33	0.10	-0.52	-0.14
Plymouth	-0.35	0.09	-0.53	-0.18
Porsche	0.02	0.11	-0.19	0.23
Renault	-0.29	0.11	-0.55	-0.08
Saab	0.08	0.1	-0.12	0.26
Subaru	-0.25	0.08	-0.42	-0.09
Toyota	-0.22	0.08	-0.37	-0.07
Volkswagen	-0.15	0.08	-0.31	0.01
Volvo	-0.1	0.09	-0.28	0.08
carheight	-0.03	6.81e-03	-0.04	-0.02
curbweight	1.43e-03	3.26e-04	7.82e-04	2.07e-03
peakrpm	-4.85e-04	3.25e-04	-1.13e-03	1.57e-04
citympg	-0.04	0.01	-0.06	-0.02
I(wheelbase^2)	1.06e-04	2.43e-05	5.75e-05	1.54e-04
I(carlength^2)	-2.17e-05	8.07e-06	-3.77e-05	-5.78e-06
I(carwidth^2)	1.98e-04	8.67e-05	2.67e-05	3.69e-04
I(curbweight^2)	-1.71e-07	6.14e-08	-2.92e-07	-4.98e-08
I(enginesize^2)	3.96-06	1.55e-06	8.94e-07	7.01e-06
I(peakrpm^2)	5.27e-08	3.08e-08	-8.19e-09	1.14e-07
I(citympg^2)	3.5e-04	2.01e-04	-4.71e-05	7.47e-04
I(highwaympg^2)	1.84e-04	9.84e-05	-10e-06	3.79e-04

Tabella 8: Intervallo di confidenza t al 95% per le stime dei parametri.

Inoltre, poiché la nostra risposta è log(prezzo delle auto), abbiamo anche costruito l'intervallo di t-confidenza al 95% per gli effetti moltiplicativi. Abbiamo ottenuto l'intervallo di confidenza degli effetti moltiplicativi da

$$exp\left(\hat{\beta}_i \pm SE_{\hat{\beta}_i} t_{\alpha/2,165}\right)$$

	$exp(\hat{\beta}_i)$	Limite Inferiore	Limite Superiore
(Intercept)	8468.52	873.67	85612.46
Aspiration Turbo	1.12	1.06	1.18
Carbody Hardtop	0.84	0.74	0.96
Carbody Hatchback	0.80	0.71	0.90
Carbody Sedan	0.86	0.76	0.97
Carbody Wagon	0.86	0.75	0.98
enginelocation rear	1.97	1.63	2.38

Audi	0.97	0.81	1.17
BMW	1.36	1.15	1.62
Buick	1.04	0.83	1.31
Chevrolet	0.83	0.67	1.02
Dodge	0.71	0.60	0.84
Honda	0.82	0.69	0.97
Isuzu	0.88	0.73	1.06
Jaguar	0.89	0.65	1.21
Mazda	0.90	0.77	1.06
Mercury	0.85	0.64	1.11
Mitsubishi	0.68	0.58	0.80
Nissan	0.85	0.73	1
Peugeot	0.72	0.59	0.87
Plymouth	0.70	0.59	0.84
Porsche	1.02	0.83	1.26
Renault	0.75	0.60	0.93
Saab	1.08	0.89	1.30
Subaru	0.78	0.66	0.91
Toyota	0.80	0.69	0.94
Volkswagen	0.86	0.73	1.02
Volvo	0.91	0.76	1.08
carheight	0.97	0.96	0.98
curbweight	1.00	1.00	1.00
peakrpm	1.00	1.00	1.00
citympg	0.96	0.94	0.99
I(wheelbase^2)	1.00	1.00	1.00
I(carlength^2)	1.00	1.00	1.00
I(carwidth^2)	1.00	1.00	1.00
I(curbweight^2)	1.00	1.00	1.00
I(enginesize^2)	1.00	1.00	1.00
I(peakrpm^2)	1.00	1.00	1.00
I(citympg^2)	1.00	1.00	1.00
I(highwaympg^2) 1.00	1.00	1.00	

Tabella 9: Intervallo di confidenza t al 95% per l'effetto moltiplicativo.

Infine, abbiamo effettuato un'analisi dei residui per verificare se i nostri residui seguono le assunzioni di normalità e varianza costante. Per la nostra analisi, abbiamo deciso di utilizzare i residui standardizzati invece dei residui per interpretazioni più semplici. Abbiamo ottenuto i residui standardizzati trovando la differenza tra i valori osservati e quelli stimati e dividendo per i valori stimati. Successivamente, abbiamo tracciato i grafici dei residui per ulteriori analisi.

Dal grafico dei residui standardizzati rispetto ai valori stimati, possiamo vedere che non c'è un pattern evidente all'interno dei residui standardizzati. Dal Normal QQ plot dei residui standardizzati, possiamo vedere che la maggior parte dei residui standardizzati ruota attorno alla linea di normalità e non ci sono residui che si discostano significativamente da quella linea. Pertanto, i risultati dei residui standardizzati indicano che il nostro modello è un buon fit.

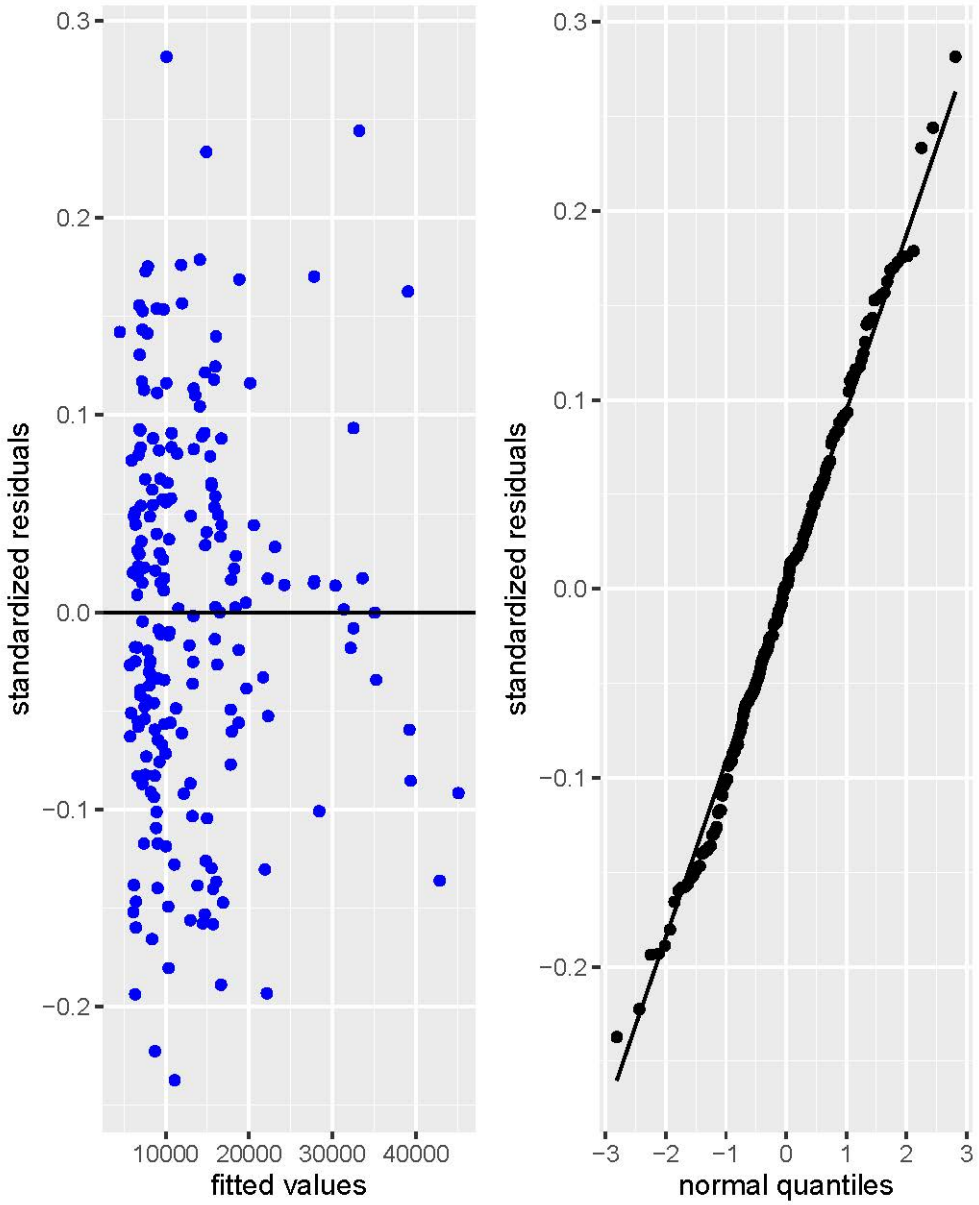

Tabella 10: Distribuzione del prezzo. (Sinistra) Residuo standard vs. Valori adattati (Destra) Grafico QQ normale dei residui standard.

CODICE

```
# modello gamma con tutte le variabili esplicative

gamma_model_price1<-glm(price~., data= car_dat3[, -c(7,9)],family= Gamma(link = "log"))

sum_mod_gamma<-summary(gamma_model_price1)

# utilizzo di stepAIC per effettuare la selezione stepwise

library(MASS)

gamma_step_model <- stepAIC(gamma_model_price1, direction = "both",

                trace = TRUE)

gamma_step_sum<-summary(gamma_step_model)

# gamma_squareb<-glm(formula = price ~ aspiration + carbody + enginelocation +

# cylindernumber + car_company + wheelbase + carlength + carwidth +

# carheight + curbweight + enginesize + boreratio + peakrpm +

# citympg + highwaympg, family = Gamma(link = "log"), data = car_dat3[,-c(7, 9)])

gamma_square<-glm(formula = price ~ fueltype + aspiration + doornumber + carbody + drivewheel + enginelocation +

            cylindernumber + car_company + symboling + wheelbase + carlength + carwidth +

            carheight + curbweight + enginesize + boreratio + stroke + compressionratio + horsepower + peakrpm +

            citympg + highwaympg+ I(wheelbase^2)+I(carlength^2)+

            I(carwidth^2)+I(carheight^2)+I(curbweight^2)+I(enginesize^2)+

            I(boreratio^2)+I(peakrpm^2)+I(citympg^2)+I(highwaympg^2)

           , family = Gamma(link = "log"), data = car_dat3[,-c(7, 9)])

# Selezione stepwise per il modello quadratico.

gamma_square_step<-stepAIC(gamma_square,direction = "both",trace=FALSE)

# lrttest per modello completo vs modello stepwise (effetto principale)
```

```
loglik_full<-logLik(gamma_model_price1)

loglik_step<-logLik(gamma_step_model)

test_stat1<-loglik_full-loglik_step

p_val1<-1-pchisq(test_stat1,8)

# Il nostro modello è uguale al modello completo (effetto principale)

# lrttest per modello quadratico completo vs modello stepwise quadratico

loglik_sq_full<-logLik(gamma_square)

loglik_sq_step<-logLik(gamma_square_step)

test_stat2<-loglik_sq_full-loglik_sq_step

pval2<-1-pchisq(test_stat2,22)

# Il nostro modello è uguale al modello completo (modello quadratico)

# Confronto tra modello ad effetto principale e modello quadratico

AIC(gamma_step_model)

AIC(gamma_square_step)

# Il modello quadratico ha l'AIC inferiore rispetto al modello ad effetto principale.
Pertanto, il modello quadratico è il migliore.

# adeguatezza del modello

deviance(gamma_step_model)

deviance(gamma_square_step)

pchisq(deviance(gamma_square_step),df.residual(gamma_square_step),lower.
tail = FALSE)

# il modello è adeguato

sum_square_step<-summary(gamma_square_step)

# Intervallo di confidenza al 95% per gli stime

coefficient_dat<-sum_square_step$coefficients[,c(1,2)]

LL<-coefficient_dat[,1]-
qt(0.975,df.residual(gamma_square_step))*coefficient_dat[,2]
```

```
UL<-
coefficient_dat[,1]+qt(0.975,df.residual(gamma_square_step))*coefficient_dat[
,2]

conf_int_dat<-data.frame(coefficient_dat,LL,UL)

names(conf_int_dat)[3]<-"Lower Limit"

names(conf_int_dat)[4]<-"Upper Limit" (conf_int_dat)

# Pseudo R2 per il modello stepwise quadratico.

pseudo_R2_square_step<-1-
(sum_square_step$deviance/sum_square_step$null.deviance)

#0.9661938

# test con residui standardizzati

predict_car<-fitted(gamma_square_step)

price_actual<-car_dat3$price

std_res_car<-(price_actual-predict_car)/predict_car

# grafico dei residui standardizzati

pdf("C:/Users/Juli/Desktop/Progetto/Figures/std_residui_pattern.pdf", width =
8, height = 5)

plot(predict_car, std_res_car, xlab = 'valori previsti', ylab = 'residui
standardizzati') abline(h = 0, col = "blue") dev.off()

# analisi: nessun pattern evidente nei residui standardizzati indicando varianza
costante

# grafico di normalità dei residui standardizzati

pdf("C:/Users/Juli/Desktop/Progetto/Figures/qqplot_std_res_auto.pdf", width
= 8, height = 5)

qqnorm(std_res_car,ylab="quantili campione",xlab="Quantili teorici")

qqline(std_res_car)

dev.off()

# analisi: tutti i residui standardizzati si trovano sulla linea qq, indicando che i
residui seguono una distribuzione normale approssimativa
```

1.4 Conclusioni

Il nostro miglior modello include le covariate: aspirazione, tipo di motore, tipo di carrozzeria, casa automobilistica, altezza dell'auto, peso a vuoto, giri al minuto massimi, consumo in città, passo, lunghezza dell'auto, larghezza dell'auto, dimensione del motore e consumo in autostrada. Prima di procedere con la conclusione di questa analisi, ecco alcuni punti salienti delle nostre interpretazioni dei parametri:

1. Il prezzo dell'auto aumenta di circa 2 volte quando un'auto ha un motore posteriore.

2. Le carrozzerie diverse dalla decappottabile (livello di riferimento) fanno diminuire il prezzo dell'auto.

3. Tra i produttori di automobili, BMW ha la migliore reputazione poiché il prezzo dell'auto aumenta di 1,4 volte quando il marchio è BMW. Al contrario, Plymouth e Dodge hanno la peggiore reputazione (il prezzo dell'auto diminuisce del 29% quando queste aziende sono i produttori).

4. La maggior parte delle covariate numeriche nel nostro modello ha una relazione quadratica con il log(prezzo).

5. Il nostro valore di Pseudo-R^2 è eccezionalmente alto (circa intorno al 96%).

Anche se il nostro modello mostra prospettive promettenti nello studio dei prezzi delle auto negli Stati Uniti, ci sono alcune limitazioni che dobbiamo affrontare.

In primo luogo, le auto elettriche, che costituiscono il 10% del mercato automobilistico attuale, non sono incluse nel nostro dataset. In secondo luogo, possiamo notare l'assenza di auto di lusso e auto sportive. Infine, i dati sono limitati per alcuni marchi come Mercury, che ha solo un campione nel nostro intero dataset.

Pertanto, si può concludere che, sebbene il nostro modello mostri risultati promettenti per lo studio dei prezzi delle auto, potrebbe non riflettere i prezzi attuali delle auto negli Stati Uniti a causa delle limitazioni dei nostri dati.

2 Classificazione Della Cardiopatia

2.1 Introduzione

L'insufficienza cardiaca, può essere definita in generale come una condizione che si verifica quando il cuore non riesce a fornire il fabbisogno di ossigeno e sangue del corpo. Secondo l'ultimo rapporto statistico annuale dell'American Heart Association e del National Institutes of Health, circa 6,2 milioni di adulti negli Stati Uniti soffrono di insufficienza cardiaca. Inoltre, nel 2018, l'insufficienza cardiaca è stata menzionata su 379.800 certificati di morte (13,4%) e ha comportato un costo annuo di circa 30 miliardi di dollari. Questo suggerisce che identificare i comportamenti di salute e i fattori di rischio principali che influenzano l'insufficienza cardiaca è fondamentale non solo per la salute della comunità, ma anche per l'economia. Pertanto, abbiamo deciso di analizzare il dataset "Heart Failure prediction" per trovare le variabili che giocano un ruolo chiave nell'insufficienza cardiaca. Poiché la risposta nel nostro dataset è binaria (0 o 1), abbiamo utilizzato la regressione logistica per modellare la probabilità di avere insufficienza cardiaca. Inoltre, abbiamo eseguito la selezione delle variabili per scegliere il miglior modello e determinare i principali fattori nell'insufficienza cardiaca (dettagli nella sezione Analisi Statistiche).

Questo studio ha generalmente rivelato fattori causali nell'insufficienza cardiaca come sesso, angina da sforzo (un tipo di dolore toracico durante l'esercizio fisico), tipi distinti di dolore toracico e livello di colesterolo al quadrato.

La generalizzazione dei nostri risultati è soggetta a certe limitazioni. Ad esempio, il nostro dataset non copre le generazioni più giovani (meno di 28 anni), il che porterà le analisi a essere orientate verso le età più avanzate. Un'altra questione non affrontata in questo studio è la mortalità dei pazienti. Questo potrebbe non sembrare discutibile a prima vista. Tuttavia, molti pazienti con dolore toracico asintomatico potrebbero aver vissuto senza problemi critici per tutta la vita, mentre pazienti con altri tipi di dolore toracico potrebbero aver affrontato situazioni devastanti. Questo potrebbe causare che i nostri risultati siano discutibili da diverse prospettive. In generale, il nostro studio ha concluso che ci

sono significativi fattori di rischio nell'insufficienza cardiaca. I capitoli seguenti sono una descrizione dettagliata delle nostre analisi, metodi e risultati.

2.2 Caratteristiche Dei Dati

I dati sono stati raccolti dal sito web di Kaggle (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction), dove è possibile accedere facilmente alla versione online dei nostri dati. I dati contengono 918 soggetti con 11 covariate e una risposta binaria (Insufficienza Cardiaca o meno). I dati non presentano valori mancanti e sono pronti per l'analisi. La Figura 4 rappresenta la distribuzione dell'età nei campioni in base al sesso e alla condizione cardiaca. Come è evidente, il numero di campioni maschili è drasticamente superiore a quello dei campioni femminili, indicando se il nostro dataset è sbilanciato o se gli uomini hanno più episodi di insufficienza cardiaca rispetto alle donne. Secondo due distinti studi indipendenti, gli uomini hanno più episodi di insufficienza cardiaca, il che è coerente con il nostro dataset. Un'altra osservazione nella Figura 6 è che l'età parte da 28 anni, mostrando che il nostro dataset non contiene generazioni più giovani, e la nostra analisi non è valida per le età più giovani.

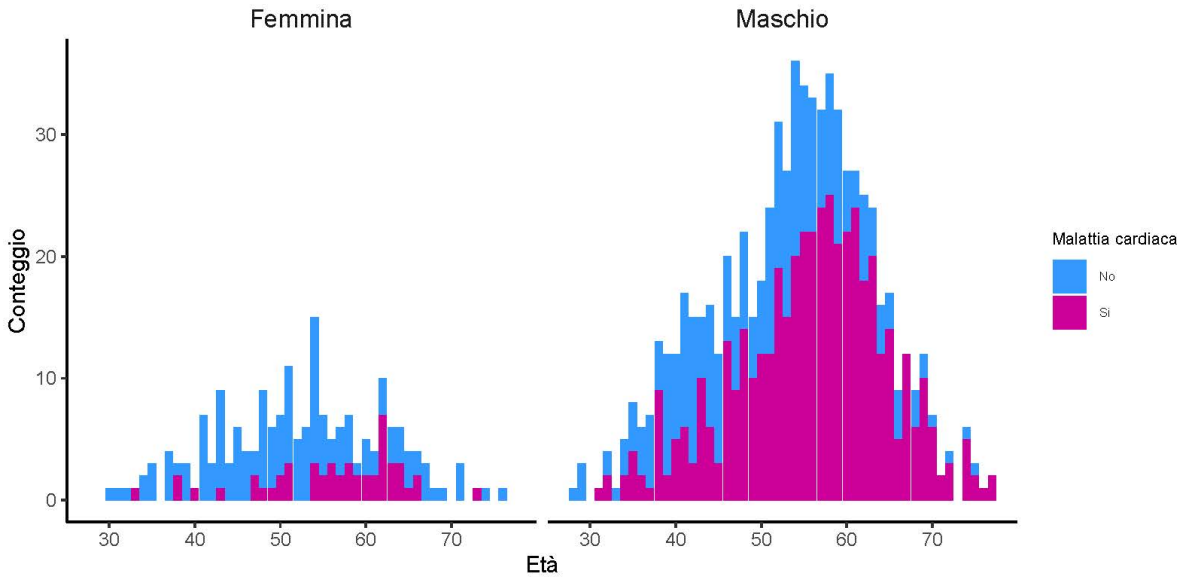

Figura 5: La distribuzione dell'età dei campioni rispetto al loro sesso e alla condizione cardiaca.

Come mostrato nella Figura 6(A), il tipo di dolore toracico asintomatico è più frequente negli uomini e nelle donne, e anche i campioni con tipo di dolore toracico asintomatico sono più esposti all'insufficienza cardiaca. Questa osservazione non è sorprendente perché i pazienti con insufficienza cardiaca possono mostrare sintomi. La Figura 6(B) presenta che i campioni con angina da sforzo sono più esposti all'insufficienza cardiaca in entrambi i gruppi di sesso.

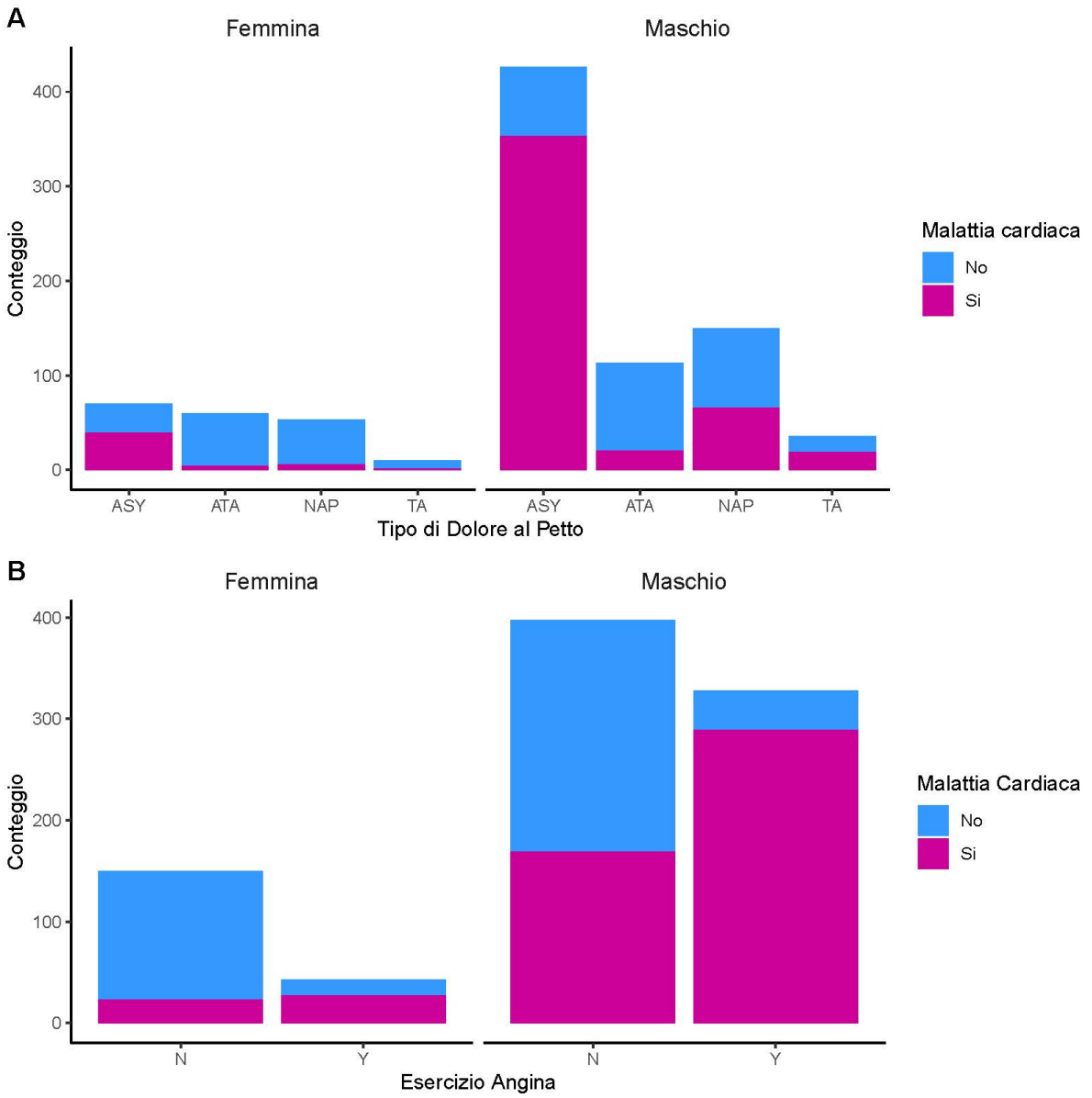

Figura 6: Il numero di campioni con diversi tipi di dolore al petto e angina da sforzo. (A) I tipi distintivi di dolore al petto nei diversi generi. TA: Angina Tipica, ATA: Angina Atipica, NAP: Dolore Non Anginoso, ASY: Asintomatico (B) Numero di campioni con angina da sforzo rispetto al loro sesso. Y: Sì, con angina da sforzo e N: No, senza angina da sforzo.

Nella Figura 7, possiamo vedere che i pazienti con problemi cardiaci hanno un depressioneST più alto e, come si potrebbe sospettare, l'insufficienza cardiaca potrebbe essere influenzata anche dal quadrata della depressioneST. Pertanto, abbiamo deciso di includere le forme quadratiche delle variabili numeriche nel nostro modello. Una descrizione dettagliata dei principali risultati, insieme alla nostra conclusione, è fornita nei capitoli successivi.

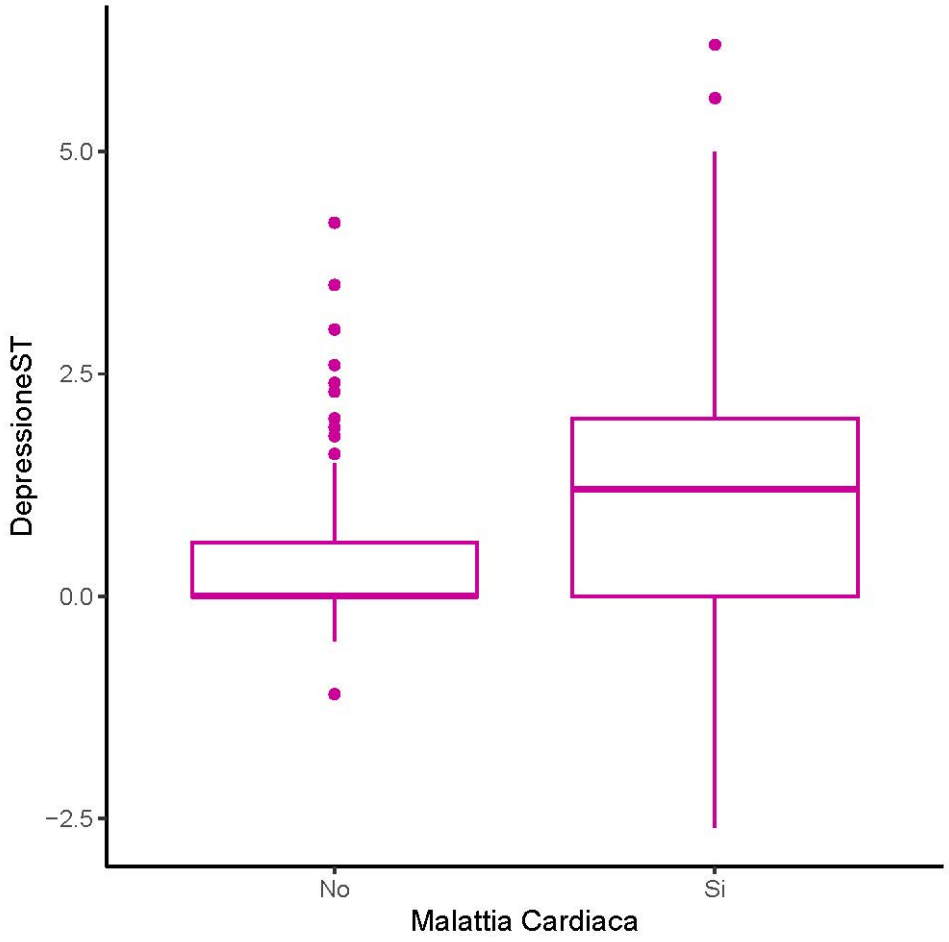

Figura 7: La relazione tra avere malattie cardiache e la depressioneST.

CODICE

#Carga dei dati

malattia_cardiaca <- read_csv("C:/Users/Juli/Desktop/Progetto/data/heart.csv")

```
colnames(malattia_cardiaca) <- c("Età", "Sesso", "TipoDoloreToracico", "PressioneSanguignaRiposo",
                                 "Colesterolo", "GlicemiaBasale", "ECGRiposo",
                                 "FrequenzaCardiacaMassima", "AnginaEsercizio", "DepressioneST",
                                 "PendenzaST", "MalattiaCardiaca")
#Re-codificare i fattori
malattia_cardiaca$MalattiaCardiaca <- factor(malattia_cardiaca$MalattiaCardiaca, levels = c(0, 1), labels = c('No', 'Si'))
malattia_cardiaca$Sesso <- factor(malattia_cardiaca$Sesso, levels = c('F', 'M'), labels = c('Femmina', 'Maschio'))
# Grafico del tipo di dolore toracico per sesso
df <- malattia_cardiaca %>% group_by(MalattiaCardiaca, TipoDoloreToracico, Sesso) %>% count()
p_tipo_dolore <-
 ggplot(df, aes(y = n, x=TipoDoloreToracico, fill=MalattiaCardiaca)) + facet_grid(.~Sesso)+
 geom_bar(stat = 'identity') +
 xlab('Tipo di Dolore al Petto') + ylab('Conteggio')+
 scale_fill_manual(values=c("#3399ff", "#cc0099"),
                   name = "Malattia cardiaca")+
 theme_classic() +
 theme(strip.background = element_blank(), strip.text = element_text(size = 12))
p<-
```

```
 ggplot(malattia_cardiaca, aes(x=Età, fill = MalattiaCardiaca))+ facet_grid(.~Sesso)+
 geom_bar() +
 xlab('Età') + ylab('Conteggio')+
 scale_fill_manual(values=c("#3399ff", "#cc0099"),
                   name = "Malattia cardiaca")+
 theme_classic() +
 theme(line = element_line(size = 0.5), strip.background = element_blank(), strip.text = element_text(size = 12),
       legend.title = element_text(size=8), legend.text = element_text(size = 6)) +
 guides(color = guide_legend(override.aes = list(size = 0.2)))
p
ggsave('C:/Users/Juli/Desktop/Progetto/Figures/Età_heart_Sesso.pdf', p, width = 8, height = 4)


##Box plot Colesterolo
p <- ggplot(malattia_cardiaca, aes(y = Colesterolo, x=MalattiaCardiaca))+
 geom_boxplot(colour = "#cc0099") +
 xlab('Malattia Cardiaca') + ylab('Colesterolo')+
 scale_fill_manual(values=c("#3399ff", "#cc0099"),
                   name = "Malattia Cardiaca")+
 theme_classic() +
```

```
 theme(strip.background = element_blank(), strip.text = element_text(size = 12))

p

ggsave('C:/Users/Juli/Desktop/Progetto/Figures/colesterolo.jpeg', p, width = 4, height = 5)


### exercise

df <- malattia_cardiaca %>% group_by(MalattiaCardiaca, AnginaEsercizio, Sesso) %>% count()

ex_ang_plot <-

 ggplot(df, aes(y = n, x=AnginaEsercizio, fill=MalattiaCardiaca)) + facet_grid(.~Sesso)+

 geom_bar(stat = 'identity') +

 xlab('Esercizio Angina') + ylab('Conteggio')+

 scale_fill_manual(values=c("#3399ff", "#cc0099"),

                   name = "Malattia Cardiaca")+

 theme_classic() +

 theme(strip.background = element_blank(), strip.text = element_text(size = 12))

p <-

 ggpubr::ggarrange(p_tipo_dolore, ex_ang_plot,

                   labels = c("A", "B"),

                   ncol = 1, nrow = 2)


p
```

```
ggsave('C:/Users/Juli/Desktop/Progetto/Figures/tipo_dolore_exercise_angina.pdf', p, width = 8, height = 8)

p <-

 ggplot(malattia_cardiaca, aes(y = DepressioneST, x=MalattiaCardiaca))+

 geom_boxplot(colour = "#cc0099") +

 xlab('Malattia Cardiaca') + ylab('DepressioneST')+

 theme_classic() +

 theme(strip.background = element_blank(), strip.text = element_text(size = 12))

p

ggsave('C:/Users/Juli/Desktop/Progetto/Figures/Depressione.pdf', p, width = 5, height = 5)
```

2.3 Analisi Statistica

Per cominciare, abbiamo assegnato a ciascuna caratteristica un nome di variabile.

Nome Variabile	Definizione	Spiegazione	Variabile
Età	Età del paziente	Anni	x_1
Sesso	Sesso del paziente	M: Maschio, F: Femmina	x_2
TipoDoloreToracico	Tipo di dolore al petto	TA: Tipico Angina, ATA: Atipico Angina, NAP: Dolore non da Angina, ASY: Asintomatico	x_3
PressioneSanguignaRiposo	Pressione sanguigna a riposo	mm Hg	x_4
Colesterolo	Colesterolo	1: se a digiuno > 120 mg/dl, 0: altrimenti	x_5

GlicemiaBasale	Livello di zucchero nel sangue a digiuno	Livello del sangue a digiuno [1: se a digiuno > 120 mg/dl, 0: altrimenti]	x_6
ECGRiposo	Risultato elettrocardiogramma a riposo	Normal: Normale, ST: onde ST-T anormali (T wave inversione e/o ST of > 0.05 mV), LVH: segnalare ipertrofia ventricolare	x_7
FrequenzaCardiacaMassima	Massima frequenza cardiaca registrata	Valore numericofra 60 e 202	x_8
AnginaEsercizio	Angina indotta da esercizio	Y: Si, N: No	x_9
DepressioneST	Depressione = ST	Depressione misurata con valori numerici	x_{10}
PendenzaST	La pendenza massima dell esercizio ST	Up: pendanza in salita, Flat: piano, Down: pendenza in discesa	x_{11}
MalattiaCardiaca	Classe di output	1: Malattia Cardiaca, 0: Normal	y

Tabella 11: Dizionario dei dati.

Modellazione di tutte le covariate principali

Poiché la risposta è avere o meno una malattia cardiaca, si tratta di una risposta distribuita binomialmente. Pertanto, abbiamo utilizzato un modello di regressione logistica. Inizialmente, abbiamo creato un modello logistico con tutti gli effetti principali, e la formula del modello è:

$$log\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11} \beta_i x_i$$

dove p è la probabilità di avere la malattia cardiaca, e $\beta_i, i = 1, 2, \dots, 11$ sono i coefficienti dei parametri, β_0 è l'intercetta. Abbiamo quindi adattato il modello e ottenuto i parametri stimati mostrati nella tabella sottostante.

	Stima	Pr(>|z|)	Significativitò
(Intercept)	-1.16	0.411	not significant
Età	0.0166	0.21	not significant
SessoM	1.47	1.6e-07	***
DoloreToracicoATA	-1.83	2.03e-08	***
DoloreToracicoNAP	-1.69	2.34e-10	***
DoloreToracicoTA	-1.49	0.00058	**

PressioneSanguignaRiposo	0.00419	0.485	not significant
Colesterolo	-0.00411	0.000154	**
GlicemiaBasale	1.14	3.59e-05	***
ECGRiposoNormale	-0.177	0.515	not significant
ECGRiposo	-0.269	0.443	not significant
FrequenzaCardiacaMassima	-0.00429	0.393	not significant
AnginaEsercizioY	0.9	0.000231	**
Depressione	0.381	0.00131	*
PendenzaPianaST	1.45	0.000703	**
PendenzaSalitaST	-0.994	0.0272	.

Tabella 12: Stima e significatività degli effetti del modello completo.
*Legenda: eccessivo *** forte ** moderato * borderline.*

Dalla tabella, possiamo vedere che alcune variabili non sono significative. Dovremmo eliminare alcune variabili per rendere il modello più semplice. Abbiamo scelto di fare una selezione stepwise backward. I parametri stimati sono mostrati nella tabella sottostante.

	Stime	Pr(>|z|)	Significatività
(Intercept)	-1.72	0.0436	.
Età	0.0231	0.0518	not significant
SessoM	1.47	1.36e-07	***
DoloreToracicoATA	-1.86	8.89e-09	***
DoloreToracicoNAP	-1.72	6.13e-11	***
DoloreToracicoTA	-1.49	0.000494	**
Colesterolo	-0.00398	0.000106	**
GlicemiaBasale	1.13	3.41e-05	***
AnginaEsercizioY	0.936	8.21e-05	***
Depressione	0.377	0.00121	*
PendenzaPianaST	1.46	0.000654	**
PendenzaSalitaST	-1.03	0.0211	.

Tabella 13: Stima e significatività degli effetti del modello ridotto.
*Legenda: eccessivo *** forte ** moderato * borderline.*

Dalla tabella, possiamo vedere che ora tutte le variabili sono significative. La formula per il primo modello è:

$$log\frac{\hat{p}}{1-\hat{p}} = -1.7 + 0.023Age + 1.5SessoM - 1.9DoloreAlPettoTipoATA$$
$$- 1.7DoloreAlPettoTipoNAP - 1.5DoloreAlPettoTipoTA$$
$$- 0.0040Colesterolo - 1.1DigiunoBS$$
$$+ 0.94EsercizioAnginaY + 0.38OldPeak + 1.5\text{ST\_slopeFlat}$$
$$- 1.0\text{ST\_slopeUp}$$

Abbiamo testato se il modello selezionato è sufficientemente buono per rappresentare il modello originale. Abbiamo eseguito un test di rapporto di verosimiglianza per le variabili selezionate stepwise per vedere se l'eliminazione è giustificata. L'ipotesi nulla è

$$H_0: \beta_{\text{RestingBP}} = \beta_{\text{RestingECG}} = \beta_{\text{MaxHR}} = 0$$

$$vs$$

$$H_1 = Almeno\ uno\ dei\ parametri\ diverso\ da\ 0$$

Abbiamo usato la formula per ottenere la statistica LLR come:

$$LLR = 2\big(\ell(Modello\ Completo) - \ell(Modello\ Ridotto)\big)$$
$$= 2 * (-297.0925 + 297.9042) = 0.804$$

con gradi di libertà pari a 4. Il valore p calcolato è 0.8046016, che è molto alto. Pertanto, non possiamo rifiutare H_0. Possiamo quindi accettare il modello ridotto.

Modellazione del quadrato delle variabili numeriche

Abbiamo quindi esaminato il quadrato delle variabili numeriche. Abbiamo adottato questo approccio perché la risposta potrebbe avere un effetto quadratico delle variabili numeriche, mentre il quadrato delle variabili categoriche non fa alcuna differenza. Il modello delle probabilità è:

$$log\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11}\beta_i x_i + \beta_{12}x_1^2 + \beta_{42}x_4^2 + \beta_{72}x_7^2 + \beta_{102}x_{10}^2$$

Il modello è stato adattato usando R. Tutti i parametri e la loro significatività sono mostrati nella tabella sottostante.

	Stime	Pr(>|z|)	Significatività
Età	-0.029	0.786	not significant
SessoM	1.49	2.28e-07	***

DoloreToracicoATA	-1.71	2.21e-07	***
DoloreToracicoNAP	-1.68	8.06e-10	***
DoloreToracicoTA	-1.38	0.00231	*
PressioneSanguignaRiposo	-0.0286	0.547	not significant
Colesterolo	-0.0116	2.87e-05	***
GlicemiaBasale	1.08	0.000173	**
ECGRipostoNormale	-0.165	0.553	not significant
ECGRiposo	-0.296	0.41	not significant
FrequenzaCardiacaMassima	-0.0229	0.581	not significant
AnginaEsercizioY	1.08	2.09e-05	***
Depressione	-0.304	0.333	not significant
PendenzaPianaST	1.77	0.000112	**
PendenzaSalitaST	-7.29e-01	1.26e-01	not significant
I(MaxHR^2)	7.45e-05	6.24e-01	not significant
I(Età^2)	4.72e-04	6.37e-01	not significant
I(Colesterolo^2)	2.67e-01	1.96e-02	.
I(GlicemiaBasale^2)	1.20e-04	4.95e-01	not significant
I(Colesterolo^2)	2.17e-05	2.22e-03	*

Tabella 14: Stima e significatività del modello con termini al quadrato.
*Legenda: eccessivo *** forte ** moderato * borderline.*

Tuttavia, ci sono ancora alcune covariate inutili nel modello. Per semplificare il modello, è stata utilizzata una selezione stepwise backward per il quadrato delle variabili numeriche. I parametri e la loro significatività del modello selezionato sono mostrati nella tabella sottostante.

	Stime	Pr(>|z|)	Significatività
(Intercept)	-1.05e+00	1.25e-01	not significant
SessoM	1.50e+00	1.00e-07	***
DoloreToracicoATA	-1.72e+00	1.00e-07	***
DoloreToracicoNAP	-1.68e+00	0.00e+00	***
DoloreToracicoTA	-1.37e+00	1.98e-03	*
Colesterolo	-1.18e-02	9.90e-06	***
GlicemiaBasale	1.05e+00	2.10e-04	**
AnginaEsercizioY	1.02e+00	1.92e-05	***
PendenzaPianaST	1.74e+00	1.20e-04	**

PendenzaSalitaST	-7.23e-01	1.22e-01	not significant
I(Età^2)	2.22e-04	4.77e-02	.
I(Depressione^}2)	1.70e-01	2.21e-04	**
I(Colesterolo^2)	2.24e-05	1.44e-03	*

*Tabella 15: Stima e significatività del modello ridotto con termini al quadrato. Legenda: eccessivo *** forte ** moderato * borderline.*

Abbiamo quindi applicato il test di rapporto di verosimiglianza per ottenere:

$$LLR = 2\big(\ell(Modello\ Completo) - \ell(Modello\ Ridotto)\big)$$
$$= 2 * (-286.9175 + 288.5404) = 3.245983$$

con 4 gradi di libertà. Il valore p è 0.9179859, il che significa che la riduzione è altamente probabile. La formula per il secondo modello è:

$$log\frac{\hat{p}}{1-\hat{p}} = -1.1 + 22 * 10^{-4}Age^2 + 1.5SessoM$$
$$- 1.7DoloreAlPettoTipoATA - 1.7DoloreAlPettoTipoNAP$$
$$- 1.4DoloreAlPettoTipoTA - 0.0012Colesterolo$$
$$+ 1.0DigiunoBS + 1EsercizioAnginaY + 1.7OldPeak^2$$
$$+ 1.7\mathrm{ST}_{\mathrm{slopeFlat}} - 7.2\mathrm{ST}_{\mathrm{slopeUp}} + 2.2 * 10^{-5}Colesterolo^2$$

Modellazione dell'interazione delle variabili numeriche

Abbiamo anche modellato l'interazione tra i termini numerici, e il modello è:

$$log\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11}\beta_i x_i + \sum_{i\in\mathcal{N}}\beta_{i2}x_i^2 + \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\beta_{ij}x_i x_j$$

Dove $\mathcal{N} = \{1,4,7,10\}$ Abbiamo utilizzato la selezione stepwise backward per ridurre il modello e abbiamo ottenuto tutti i coefficienti mostrati nella tabella sottostante.

	Estimate	Pr(>|z|)	significance
(Intercept)	-1.95	0.129	not significant
Età	0.0213	0.0917	.
SessoM	1.42	4.16e-07	***
DoloreToracicoATA	-1.81	2.24e-08	***
DoloreToracicoNAP	-1.66	6.4e-10	***
DoloreToracicoTA	-1.45	0.000997	***

Colesterolo	-0.00386	0.000271	***
GlicemiaBasale	1.2	1.44e-05	***
FrequenzaCardiacaMassima	0.000751	0.897	not significant
AnginaEsercizioY	0.993	6.6e-05	***
Depressione	0.668	0.373	not significant
PendenzaPianaST	1.810000	7.59e-05	***
PendenzaSalitaST	-0.732000	1.20e-01	not significant
I(Oldpeak^2)	0.324000	6.27e-03	**
FrequenzaCardiacaMassima:Despressione	-0.007750	1.42e-01	not significant

Tabella 16: Stima e significatività del modello ridotto con interazioni.
*Legenda: eccessivo *** forte ** moderato * borderline.*

Analogamente, abbiamo eseguito un test di rapporto di verosimiglianza per la riduzione, ottenendo:

$$LLR = 2\big(\ell(Modello\ Completo) - \ell(Modello\ Ridotto)\big)$$
$$= 2 * (-288.2246 + 292.3055) = 8.161785$$

con 11 gradi di libertà. Il valore p è 0.699, il che significa che la riduzione è accettabile. La formula per il terzo modello è:

$$log\frac{\hat{p}}{1-\hat{p}} = -1.9 + 0.02Age + 1.4SessoM - 1.8DoloreAlPettoTipoATA$$
$$- 1.7DoloreAlPettoTipoNAP - 1.4DoloreAlPettoTipoTA$$
$$- 0.0039Colesterolo + 1.19DigiunoBS + 7.5 * 10^{-4}MaxHR$$
$$+ 1EsercizioAnginaY + 0.32OldPeak^2 + 1.8\mathrm{ST}_{\mathrm{slopeFlat}}$$
$$- 0.73\mathrm{ST}_{\mathrm{slopeUp}} - 7.7 * 10^{-4}MaxHR * Oldpeak$$

Analisi dei Tre Modelli

Curve ROC

Le curve ROC (Receiver Operating Characteristics) sono utili per organizzare i classificatori e visualizzare le loro prestazioni. Le curve ROC mostrano il tasso di veri positivi sull'asse Y e il tasso di falsi positivi sull'asse X con diversi criteri decisionali. Se la curva ROC può raggiungere l'angolo in alto a sinistra nel grafico, il tasso di falsi positivi raggiunge zero e il tasso di veri positivi è 1, il che significa che il modello ha una classificazione perfetta. Pertanto, più la curva ROC si

avvicina all'angolo in alto a destra, migliori sono le prestazioni del modello. Le curve ROC per i tre modelli sono mostrate nella Figura 5.

Nel grafico ROC, le curve dei tre modelli quasi coincidono, il che significa che i tre modelli hanno prestazioni simili. Per approfondire, è stata calcolata l'area sotto la curva ROC (AUC) per confrontare i tre modelli. L'AUC per il primo modello è 0.932, per il secondo e terzo modello sono rispettivamente 0.9368 e 0.936. Le tre AUC sono molto simili, come previsto.

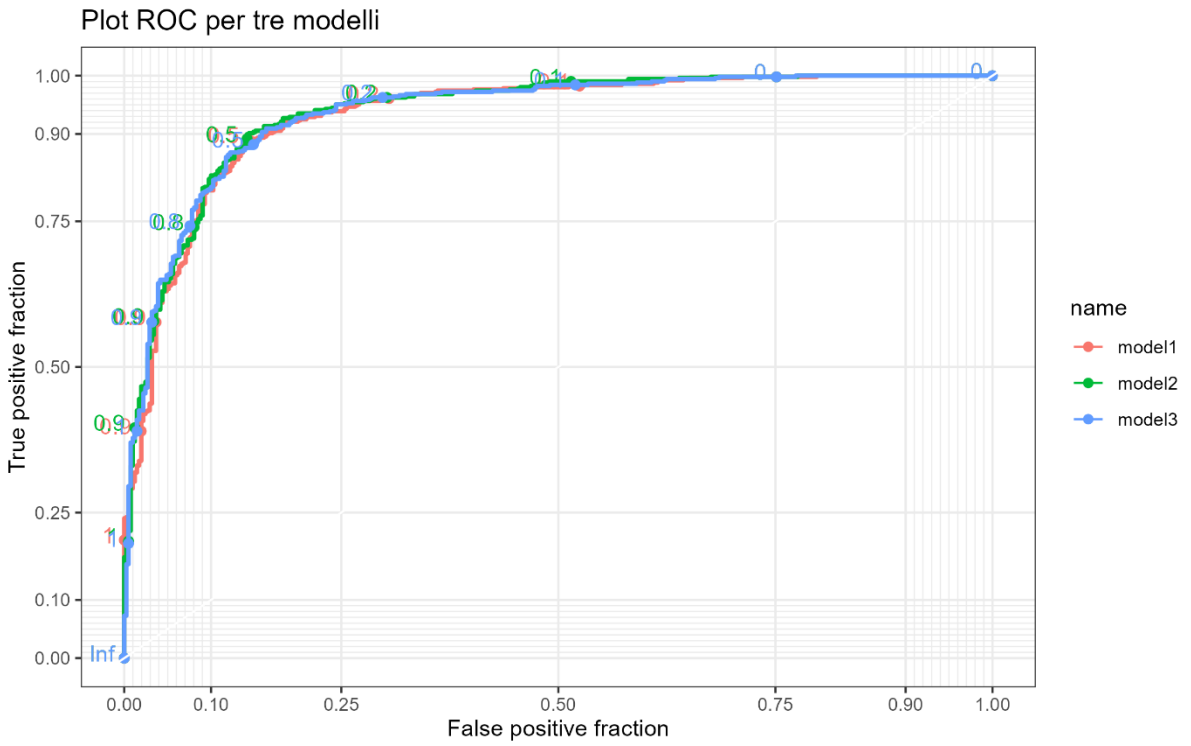

Figura 8: Le curve ROC per i tre modelli.

Pseudo-R quadro

Il Pseudo-R quadro a volte viene utilizzato per misurare la capacità esplicativa del modello. Il Pseudo-R quadro di McFadden è definito come:

$$R^2_{MF} = 1 - \frac{LL(Full)}{LL(Null)}$$

Il Pseudo-R quadro di McFadden è compreso tra 0 e 1. Più alto è il valore del Pseudo-R quadro, maggiore è la capacità esplicativa del modello. I valori del Pseudo-R quadro di McFadden per i tre modelli sono calcolati come 0.537, 0.543 e 0.528, tra i quali il secondo modello ha il valore più alto.

Devianza Residuale

Infine, abbiamo calcolato la devianza residua per i tre modelli. L'ipotesi nulla è H_0 il modello si adatta bene contro H_1 il modello si adatta male.

Il primo modello ha una devianza residua di 595.81 con 906 gradi di libertà. Quindi il p-valore è vicino a 1, il che significa che il primo modello è un buon adattamento. Il secondo modello ha una devianza di 577.08 con 905 gradi di libertà. Quindi il p-valore è approssimativamente 1. Il terzo modello ha una devianza di 584.61 con 903 gradi di libertà e un p-valore di 1. Tra le devianze dei tre modelli, il secondo modello ha il valore più basso.

Criterio di Informazione di Akaike (AIC)

L'AIC è un metodo popolare per confrontare l'adeguatezza di più modelli, possibilmente non annidati. L'obiettivo della selezione del modello AIC è stimare la perdita di informazione quando la distribuzione di probabilità f associata al modello vero (generatore) è approssimata dalla distribuzione di probabilità g, associata al modello che deve essere valutato. Per scegliere un modello da una sequenza di candidati $M_i, i = 1, 2, \dots, K$. L'AIC è definito come:

$$AIC_i = -2logL_i + 2V_i$$

dove L_i, la massima verosimiglianza per il modello candidato i, è determinata regolando i V_i parametri liberi in modo da massimizzare la probabilità che il modello candidato abbia generato i dati osservati. Akaike ha dimostrato che il modello con la minima perdita di informazione ha il valore più basso di AIC.

L'AIC per i tre modelli sono rispettivamente 619.81, 603.08 e 614.61, tra i quali il secondo modello ha il valore più basso.

CODICE

```{r}
malattia_logit <- glm(MalattiaCardiaca ~ ., data = malattia_cardiaca, family = binomial)
summary(malattia_logit)
```

```{r}
malattia_logit_ridotto = step(malattia_logit)
```

```{r}
```

summary(malattia_logit_ridotto)

```

```{r}

LLR = 2 * (logLik(malattia_logit) - logLik(malattia_logit_ridotto))

cat(logLik(malattia_logit), logLik(malattia_logit_ridotto))

df = malattia_logit_ridotto$df.residual - malattia_logit$df.residual

p_value = 1 - pchisq(LLR, df)

print(df)

cat('La statistica del test del rapporto di verosimiglianza è', LLR, 'e il p-value è ', p_value, 'con df', df)

```

```{r}

malattia_logit_tutto_quadrato = glm(MalattiaCardiaca ~.  + I(FrequenzaCardiacaMassima^2) + I(Età^2) + I(DepressioneST^2) + I(PressioneSanguignaRiposo^2) + I(Colesterolo^2), family = binomial, data = malattia_cardiaca)

summary(malattia_logit_tutto_quadrato)

```

```{r}

malattia_logit_quadrato_ridotto = step(malattia_logit_tutto_quadrato)

summary(malattia_logit_quadrato_ridotto)

```

```{r}

LLR = 2 * (logLik(malattia_logit_tutto_quadrato) - logLik(malattia_logit_quadrato_ridotto))

df = malattia_logit_quadrato_ridotto$df.residual - malattia_logit_tutto_quadrato$df.residual

p_value = 1 - pchisq(LLR, df)

print(logLik(malattia_logit_tutto_quadrato))

print(logLik(malattia_logit_quadrato_ridotto))

cat('La statistica del test del rapporto di verosimiglianza è', LLR, 'e il p-value è ', p_value)

```

```{r}

malattia_logit_tutte_interazioni = glm(MalattiaCardiaca ~. +(FrequenzaCardiacaMassima + Età + DepressioneST + PressioneSanguignaRiposo)^2 + I(FrequenzaCardiacaMassima^2) + I(Età^2) + I(DepressioneST^2) + I(PressioneSanguignaRiposo^2), family = binomial, data = malattia_cardiaca)

summary(malattia_logit_tutte_interazioni)

```

```{r}

malattia_logit_interazioni_ridotto = step(malattia_logit_tutte_interazioni, direction = "both")

```

```{r}

summary(malattia_logit_interazioni_ridotto)

```

```{r}

malattia_logit_interazioni_ridotto = glm(formula = MalattiaCardiaca ~ Età + Sesso +
TipoDoloreToracico + Colesterolo +

  GlicemiaBasale + FrequenzaCardiacaMassima + AnginaEsercizio + DepressioneST + PendenzaST +

  I(DepressioneST^2) + FrequenzaCardiacaMassima:DepressioneST, family = binomial, data =
malattia_cardiaca)

summary(malattia_logit_interazioni_ridotto)

```

```{r}

LLR = 2 * (logLik(malattia_logit_tutte_interazioni) - logLik(malattia_logit_interazioni_ridotto))

df = malattia_logit_interazioni_ridotto$df.residual - malattia_logit_tutte_interazioni$df.residual

p_value = 1 - pchisq(LLR, df)

print(logLik(malattia_logit_tutte_interazioni))

print(logLik(malattia_logit_interazioni_ridotto))

print(df)

cat('La statistica del test del rapporto di verosimiglianza è', LLR, 'e il p-value è ', p_value)

```


```{r}

# Modello null per il calcolo del pseudo R-quadro

null_model = glm(MalattiaCardiaca ~ 1, data = malattia_cardiaca, family = binomial)


# Pseudo R-quadro per malattia_logit_quadrato_ridotto

```
Pse_R2 = (deviance(null_model) - deviance(malattia_logit_quadrato_ridotto)) /
deviance(null_model)

cat('Pseudo-R quadro:', Pse_R2, '\n')
```

```{r}
null_model = glm(MalattiaCardiaca ~ 1, data = malattia_cardiaca, family = binomial)

# Pseudo R-quadro per malattia_logit_interazioni_ridotto

Pse_R2 = (deviance(null_model) - deviance(malattia_logit_interazioni_ridotto)) /
deviance(null_model)

cat('Pseudo-R quadro:', Pse_R2, '\n')
```

```{r}
null_model = glm(MalattiaCardiaca~ 1, data = malattia_cardiaca, family = binomial)

# Pseudo R-quadro per malattia_logit_ridotto

Pse_R2 = (deviance(null_model) - deviance(malattia_logit_ridotto)) / deviance(null_model)

cat('Pseudo-R quadro:', Pse_R2, '\n')
```

```{r}
# Curve ROC per i tre modelli

library(pROC)

invisible(plot(roc(malattia_cardiaca$MalattiaCardiaca,
                   fitted(malattia_logit_ridotto)),
               col = "#5da492",
               main = "Curve ROC: 3 modelli",
               legend = 'Modello logistico di covariate pure'))

invisible(plot(roc(malattia_cardiaca$MalattiaCardiaca,
                   fitted(malattia_logit_quadrato_ridotto)),
               print.auc = TRUE,
               col = "#206376",
               add = TRUE))

invisible(plot(roc(malattia_cardiaca$MalattiaCardiaca,
                   fitted(malattia_logit_interazioni_ridotto)),


```

```
            col = "#bee7a3",
            add = TRUE))
color = c("#5da492","#206376","#bee7a3")
```
```{r}
# Plot delle ROC con ggplot2
library(reshape2)
library(plotROC)
ROC = data.frame(h = malattia_cardiaca$MalattiaCardiaca,
              model1 = fitted(malattia_logit_ridotto),
              model2 = fitted(malattia_logit_quadrato_ridotto),
              model3 = fitted(malattia_logit_interazioni_ridotto))
ROC$h = as.numeric(as.character(ROC$h))
longtest <- melt_roc(ROC, "h", c("model3", "model2", "model1"))
p <-ggplot(longtest, aes(d = D, m = M, color = name), main = "Curva ROC per 3 modelli",
      xlab = "False Positive Rate (1-Specificity)", ylab = "True Positive Rate (Sensitivity)") +
   geom_roc() +
   style_roc() +
   ggtitle('Plot ROC per tre modelli')
p
ggsave('C:/Users/Juli/Desktop/Progetto/Figures/roc_plot.png', p, width = 8, height = 5)
```
Devizione test
```{r}
s = summary(malattia_logit_quadrato_ridotto)
dev = deviance(malattia_logit_quadrato_ridotto)
p_value = 1-pchisq(dev, s$df.residual)
cat('La devianza del modello è:', dev, 'con gradi di libertà di', s$df.residual, '\n')
cat('p_value è:', p_value, '\n')
```
```{r}
```

```
s = summary(malattia_logit_interazioni_ridotto)
dev = deviance(s)
p_value = 1-pchisq(dev, s$df.residual)
cat('La devianza del modello è:', dev, 'con gradi di libertà di', s$df.residual, '\n')
cat('p_value è:', p_value, '\n')
```

```{r}
s = summary(malattia_logit_ridotto)
dev = deviance(s)
p_value = 1-pchisq(dev, s$df.residual)
cat('La devianza del modello è:', dev, 'con gradi di libertà di', s$df.residual, '\n')
cat('p_value è:', p_value, '\n')
```

2.4 Conclusioni

Il miglior modello per prevedere le malattie cardiache è il modello 2, che contiene le variabili SexM, ChestPainTypeATA, ChestPainTypeNAP, ChestPainTypeTA, Cholesterol, FastingBS, ExerciseAnginaY, ST_slopeFlat, ST_slopeUp, Age², Oldpeak², Cholesterol². La formula del modello 2 è la seguente:

$$\begin{aligned} log\frac{\hat{p}}{1-\hat{p}} = {} & -1.1 + 22 * 10^{-4}Age^2 + 1.5SessoM \\ & - 1.7DoloreAlPettoTipoATA - 1.7DoloreAlPettoTipoNAP \\ & - 1.4DoloreAlPettoTipoTA - 0.0012Colesterolo \\ & + 1.0DigiunoBS + 1EsercizioAnginaY + 1.7OldPeak^2 \\ & + 1.7\mathrm{ST}_{\mathrm{slopeFlat}} - 7.2\mathrm{ST}_{\mathrm{slopeUp}} + 2.2 * 10^{-5}Colesterolo^2 \end{aligned}$$

Possiamo trarre alcune considerazioni da questi parametri:

1. Rapporto di probabilità per il sesso: Il rapporto di probabilità di avere una malattia cardiaca per i maschi rispetto alle femmine è 4.5, con un intervallo di

confidenza del 95% (2.6, 7.8), il che significa che i maschi hanno un rischio maggiore di avere una malattia cardiaca rispetto alle femmine.

2. Angina indotta da esercizio: Il rapporto di probabilità di avere una malattia cardiaca per qualcuno con angina indotta da esercizio rispetto a qualcuno senza è 2.8, con un intervallo di confidenza del 95% (1.7, 4.4). Questo indica che l'angina indotta da esercizio è un indicatore di malattia cardiaca.

3. Tipi di dolore toracico: Qualcuno con dolore toracico di tipo non anginoso, angina tipica o angina atipica ha meno probabilità di avere una malattia cardiaca rispetto a qualcuno con dolore toracico asintomatico.

4. Glicemia a digiuno: Il rapporto di probabilità di avere una malattia cardiaca è direttamente correlato alla glicemia a digiuno. Il rapporto di probabilità aumenta di 2.85 quando i livelli di zucchero aumentano di un'unità, con un intervallo di confidenza del 95% (1.64, 4.97).

5. Pendenza del segmento ST: Se la pendenza del segmento ST durante l'esercizio è in salita, è meno probabile avere una malattia cardiaca, seguita dalla pendenza discendente. Una pendenza piatta è più correlata alla malattia cardiaca.

6. Colesterolo: Il logaritmo delle probabilità di avere una malattia cardiaca è correlato in modo quadratico al livello di colesterolo. Se il livello di colesterolo è inferiore a 27.2, il logaritmo delle probabilità di avere una malattia cardiaca è negativamente correlato a esso. Altrimenti, è positivamente correlato. Questo risultato indica che sia un livello di colesterolo troppo alto che troppo basso possono essere correlati alla malattia cardiaca.

7. Età: Il rapporto di probabilità di avere una malattia cardiaca è positivamente correlato al quadrato dell'età. Il rapporto di probabilità aumenterà di 2.2×10^{-4} quando il quadrato dell'età aumenta di un'unità.

Ci sono ancora alcune limitazioni nella nostra analisi. Nella interpretazione del parametro (3), qualcuno con un tipo di dolore toracico asintomatico ha la minore probabilità di avere una malattia cardiaca, il che è controintuitivo. Com'è possibile che qualcuno senza dolore toracico abbia meno probabilità di avere una malattia cardiaca rispetto a quelli con dolore toracico? Questo dovrebbe essere ulteriormente indagato.

CODICE

```{r}

# Coefficienti e intervalli di confidenza del modello malattia_logit_quadrato_ridotto

malattia_logit_quadrato_ridotto$coefficients

confint.default(malattia_logit_quadrato_ridotto)

```

```{r}

# Esponenti dei coefficienti e intervalli di confidenza del modello malattia_logit_quadrato_ridotto

exp(malattia_logit_quadrato_ridotto$coefficients)

exp(confint.default(malattia_logit_quadrato_ridotto))

```