

Modelli Statistici: Progetto finale

California housing price regression

GIULIO FANTUZZI

15 Febbraio 2023

INTRODUZIONE AL DATASET

Il dataset contiene informazioni sul censimento delle case effettuato in California nel 1990. Nello specifico, esse riguardano la demografia nei distretti (reddito, popolazione, nuclei familiari), l'ubicazione dei distretti (latitudine, longitudine) e informazioni generali sulle case (numero di stanze, numero di camere da letto, età della casa). Notiamo fin da subito come i dati non siano definiti a livello delle singole case, ma siano invece relativi ai diversi distretti in cui è stata suddivisa la California. Pertanto, i valori di alcune variabili dovranno essere intesi come statistiche complessive del distretto, ad esempio totali, medie, mediane...

Per prima cosa, importiamo i dati e analizziamo le dimensioni del dataset:

```
data<- read.csv("data/california_housing.csv", header=T)
dim(data)
```

```
## [1] 20640      10
```

Il dataset presenta ben 20640 record e 10 attributi. Questi ultimi sono:

```
colnames(data)

## [1] "longitude"          "latitude"           "housing_median_age"
## [4] "total_rooms"         "total_bedrooms"       "population"
## [7] "households"          "median_income"        "median_house_value"
## [10] "ocean_proximity"
```

- **longitude**: la longitudine del distretto. Essa fornisce una misura di quanto ad Ovest esso sia;
- **latitude**: la latitudine del distretto. Essa fornisce una misura di quanto a Nord esso sia;
- **housing_median_age**: il valore mediano dell'età delle case del distretto. Esso fornisce una misura (in anni) del tempo trascorso dalla data di costruzione delle abitazioni alla data del Censimento;
- **total_rooms**: il numero totale di stanze presenti nel distretto;
- **total_bedrooms**: il numero totale di camere da letto presenti nel distretto;
- **population**: il numero di abitanti nel distretto;
- **households**: il numero di nuclei familiari nel distretto;
- **median_income**: il reddito mediano degli abitanti del distretto;
- **median_house_value**: il valore mediano delle case del distretto;
- **ocean_proximity**: variabile qualitativa che misura la distanza del distretto dall'oceano;

Il dataset contiene quasi esclusivamente variabili quantitative, eccetto **ocean_proximity** (teniamo presente che in seguito dovremo codificarla in maniera opportuna). Tra tutte le variabili, quella che risulta più indicata da utilizzare come variabile risposta per il nostro modello di regressione è **median_house_value**.

Per semplicità di trattazione, andremo a rinominarla come **Y**:

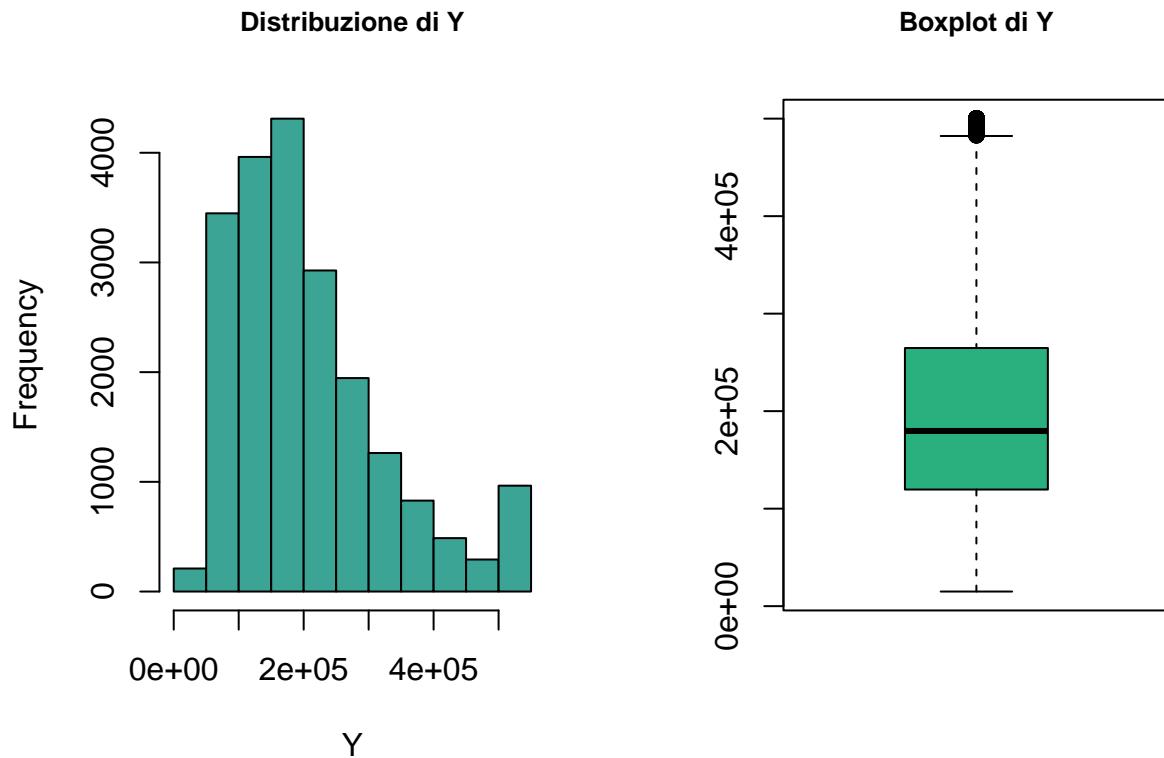
```
colnames(data)[which(colnames(data)=="median_house_value")]="Y"
```

Ancora prima di effettuare un'analisi esplorativa dei dati, è interessante soffermarsi sulla variabile risposta, evidenziandone fin da subito le caratteristiche salienti:

```
summary(data$Y)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
##  14999  119600  179700  206856  264725 500001
```

```
par(mfrow=c(1,2))
hist(data$Y, main="Distribuzione di Y",cex.main=0.8, xlab="Y", col="#3CA494")
boxplot(data$Y,main="Boxplot di Y",horizontal=F,cex.main=0.8, col="#2AAF7F")
```



Dal summary e i grafici sopra emerge una distribuzione asimmetrica della variabile risposta. Spesso, in questi casi, si applica una trasformazione alla variabile, ad esempio il logaritmo (caso particolare delle trasformazioni di Box-Cox). Dato che le trasformazioni in generale possono rendere più difficoltosa l'interpretazione del fenomeno, e visto che l'asimmetria della distribuzione di Y non è troppo accentuata rispetto alla scala di misura, ho valutato di non modificare **-per il momento-** la variabile risposta.

La distribuzione di Y nasconde anche un altro aspetto curioso: come rivela l'istogramma, la coda destra della distribuzione presenta una frequenza insolitamente alta (ciò si rispecchia anche negli outliers del boxplot).

```
sum(data$Y == max(data$Y))
```

```
## [1] 965
```

```
max(data$Y)
```

```
## [1] 500001
```

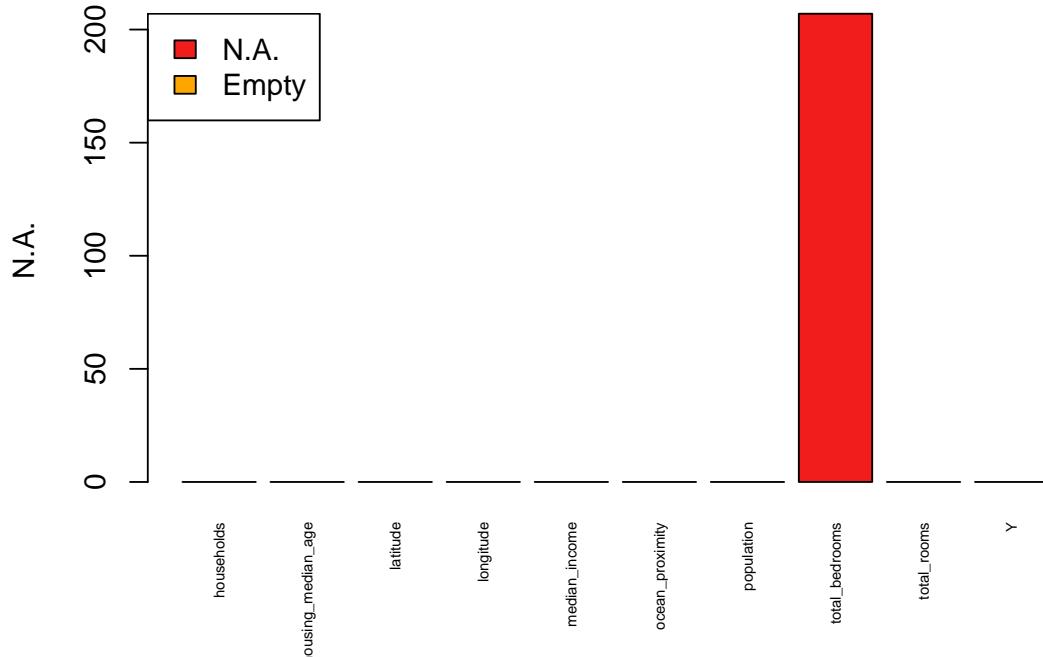
In effetti, ben 965 distretti sui 20640 totali (circa il 4,68%) presentano un prezzo mediano delle case pari a 500001\$. Ciò potrebbe essere imputabile a chi ha fornito il dataset che, probabilmente, nella sua costruzione ha applicato un effetto soglia (*threshold-effect*). Ritengo infatti che originariamente alcuni distretti presentassero valori di Y molto estremi, andando a rendere la distribuzione molto più sbilanciata. Per mantenere una situazione gestibile, pertanto, si sarà detto “Se $Y > 500000$, allora poniamo $Y = 500001$ ”.

DATA CLEANING

Prima di procedere con l'analisi esplorativa dei dati, è opportuno verificare se il nostro dataframe presenta valori non disponibili (N.A.) o campi vuoti.

```
na_values<-empty_values<- vector(mode="numeric", length=dim(data)[2])
for (i in 1: dim(data)[2]){
  na_values[i]= sum(is.na(data[,i]))
  empty_values[i]= sum(data[,i] == "", na.rm=T)
}
```

Missing Values



L'unica variabile che richiede una fase preliminare di data cleaning è **total_bedrooms** (207 N.A.). In generale, quando per una variabile quantitativa si è in presenza di valori *N.A.* si può procedere in due modi:

- 1) eliminando il record
- 2) imputando un valore

Se da una parte è bene evitare di eliminare “gratuitamente” records dal dataframe, è anche vero che dall'altra non è sempre possibile imputare un valore. Prima bisogna effettivamente capire se la mancanza del dato è significativa oppure no. In altri termini, si tratta di capire se il dato mancante è semplicemente tale, oppure se il fatto che sia mancante rappresenti un'informazione rilevante per determinare la variabile target. Ha quindi senso confrontare come si distribuisce **Y** a seconda che **total_bedrooms** abbia un valore o sia N.A.

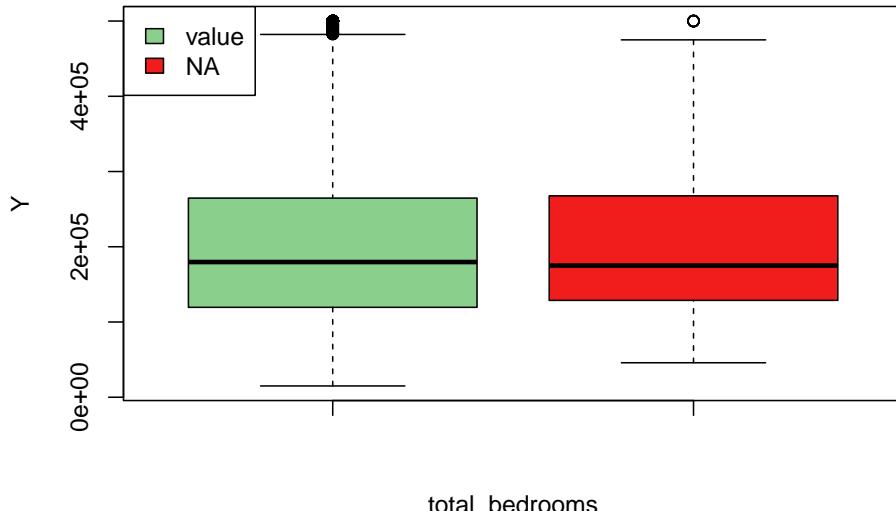
```
value<- data$Y[which(!is.na(data$total_bedrooms))]
na<- data$Y[which(is.na(data$total_bedrooms))]
summary(value)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    14999   119500  179700   206864  264700  500001

summary(na)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    45800   128750  175000   206007  267700  500001

boxplot(value,na, horizontal=F,col=c("#8ACF8B","#F11C1C"), ylab="Y", xlab="total_bedrooms")
legend("topleft",c("value", "NA"),col=c("#8ACF8B","#F11C1C"),fill=c("#8ACF8B","#F11C1C"))
```



```
var(value)/var(na)
```

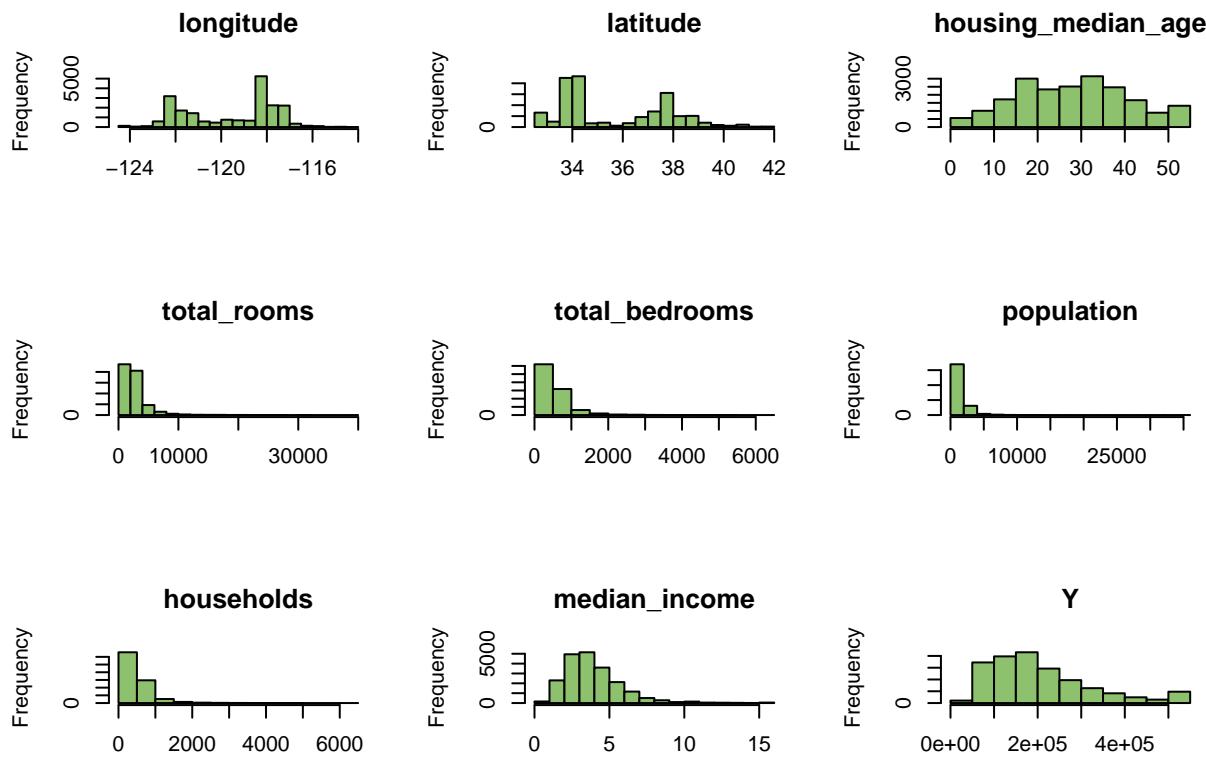
```
## [1] 1.069188
```

Il confronto tra i quantili (e le varianze) dei due gruppi ed il boxplot suggeriscono che le due distribuzioni sono molto simili, e ciò significa che i valori mancanti di **total_bedrooms** non hanno un effetto particolare sulla variabile risposta. Ora che abbiamo la conferma di questo, è possibile procedere con l'imputazione:

```
data$total_bedrooms[which(is.na(data$total_bedrooms))] <- median(data$total_bedrooms, na.rm=TRUE)
```

EDA (Exploratory Data Analysis)

Terminata la fase di data cleaning, si può procedere con un'analisi più approfondita del dataset.



DISTRIBUZIONE DELLE VARIABILI

Concentriamoci innanzitutto sulle distribuzioni delle variabili:

- **median_income** ha un'asimmetria positiva. Ciò è tanto evidente quanto prevedibile: il reddito delle persone è distribuito più o meno normalmente, ma ci sono delle persone che hanno un reddito molto più alto della media;
- **housing_median_age** ha una distribuzione più o meno uniforme (va bene!);
- **Y** come notato in precedenza è asimmetrica e rivela un intrinseco effetto soglia;
- **total_rooms**, **total_bedrooms**, **population** e **households** hanno una distribuzione fortemente asimmetrica, e assumono valori in una scala molto più ampia rispetto alle variabili presentate in precedenza. Tali variabili dovranno essere gestite in modo particolare (lo vedremo in seguito);
- **longitude** e **latitude** non presentano particolari anomalie, se non una sorta di “conca” in prossimità dei valori centrali (capiremo tra poco il motivo!);

I VALORI SONO SENSATI?

Nella fase di analisi esplorativa, oltre a valutare la distribuzione delle variabili, è essenziale capire che valori siano effettivamente contenuti nelle celle del dataframe. Infatti, comprendere in che unità sono espressi i dati che poi inseriremo nel modello è cruciale:

```
head(data, 5)
```

```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1    -122.23     37.88                 41        880          129         322
## 2    -122.22     37.86                 21       7099          1106        2401
## 3    -122.24     37.85                 52       1467           190         496
## 4    -122.25     37.85                 52       1274           235         558
## 5    -122.25     37.85                 52       1627           280         565
##   households median_income      Y ocean_proximity
## 1          126     8.3252 452600      NEAR BAY
## 2         1138     8.3014 358500      NEAR BAY
## 3          177     7.2574 352100      NEAR BAY
## 4          219     5.6431 341300      NEAR BAY
## 5          259     3.8462 342200      NEAR BAY
```

- **longitude**, **latitude** e **house_median_age** presentano valori sensati;
- **total rooms** e **total_bedrooms** sono valori rispettivamente in centinaia e migliaia. Ciò non deve stupirci, in quanto i dati sono riferiti alla totalità delle abitazioni del distretto;
- **population** inizialmente non mi era ben chiara: il dubbio è stato capire in che misura essa fosse quantificata (unità, centinaia, migliaia?). Tuttavia, sapendo che la popolazione della California nel 1990 contava circa 29,95 milioni di abitanti (*fonte Wikipedia*), è bastato fare un semplice check:

```
sum(data$population)
```

```
## [1] 29421840
```

la popolazione è quindi espressa come il totale delle persone appartenenti al distretto!

- **household** di conseguenza appare sensato
- **median_income**, invece, non sembra particolarmente sensato. Si veda il summary:

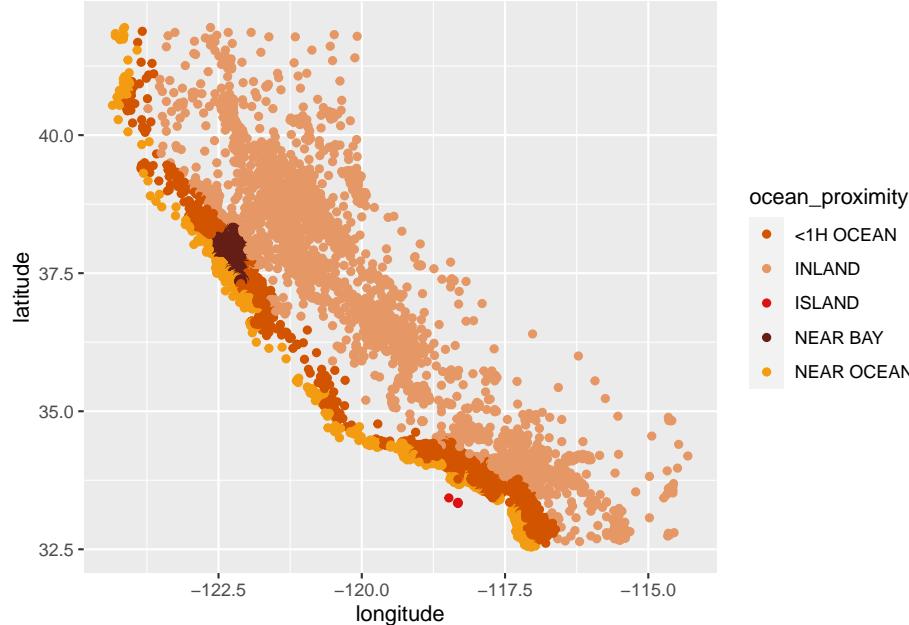
```
summary(data$median_income)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.4999  2.5634  3.5348  3.8707  4.7432 15.0001
```

Sicuramente non si può trattare di una misura in dollari. Al massimo, se i dati si riferissero al salario annuale, potrebbe aver senso considerare questa variabile in decine di migliaia di dollari.

ocean_proximity

Fino a questo momento abbiamo analizzato solamente le variabili esplicative quantitative, ma il dataset contiene anche una variabile qualitativa: **ocean_proximity**. Essa presenta in tutto 5 categorie (*si veda il grafico sotto*), che forniscono una misura della posizione geografica del distretto.



Nota: come avevamo notato in precedenza, l'istogramma di **latitude** e **longitude** presentava una sorta di “conca”, ma ciò non era di facile interpretazione. Il grafico qui sopra invece, combinando le due variabili, fornisce una visualizzazione molto più efficace: la conca dell'istogramma diventa ora una zona senza punti sulla cartina. A primo impatto potremmo pensare che una discreta parte della California all'epoca non fosse stata censita, e ciò andrebbe sicuramente a compromettere il dataset a livello di bias. In realtà la mancanza di punti deriva principalmente dalla pericolosità sismica tipica di quella zona (faglia di Sant'Andreas).

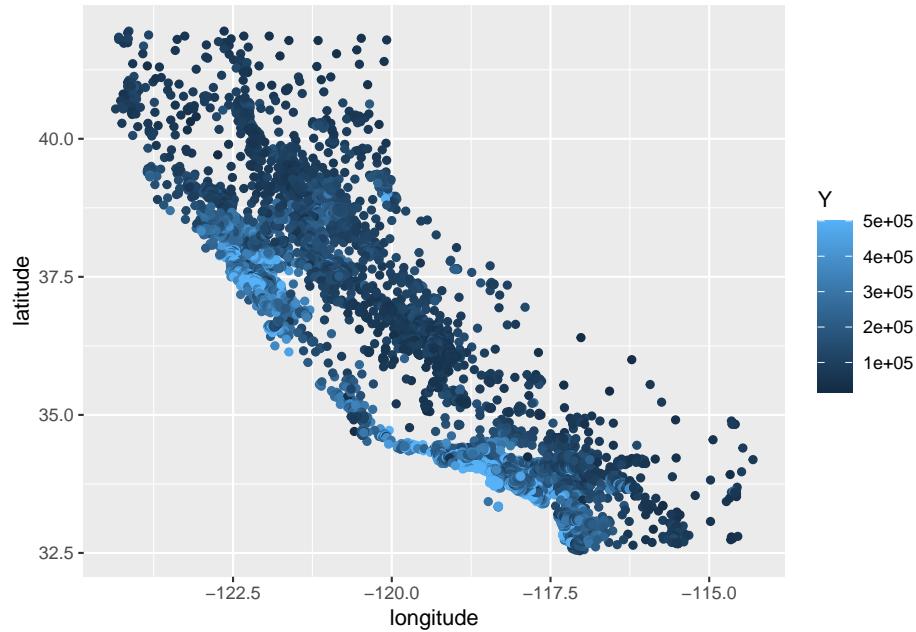
Inoltre, appare evidente la relazione tra **latitude**, **longitude** e **ocean_proximity**. Nel momento in cui andremo a costruire il modello sarà bene tenerne conto, al fine di evitare multicollinearità.

```
table(data$ocean_proximity)
```

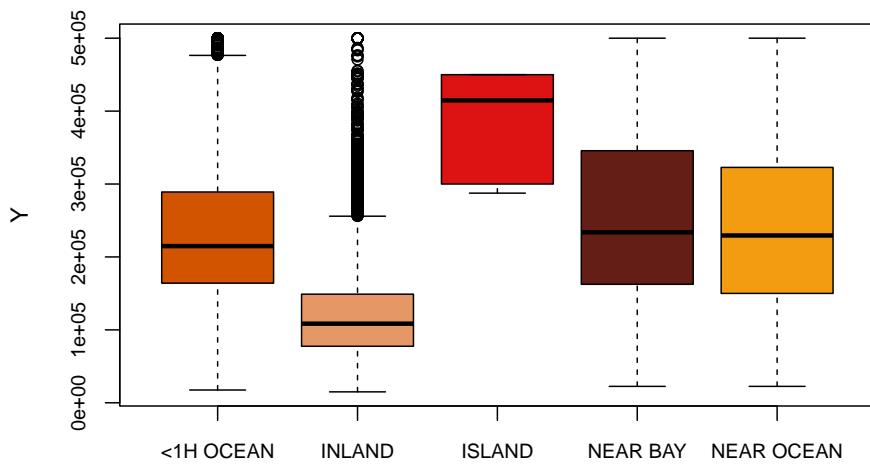
```
##          <1H OCEAN      INLAND      ISLAND    NEAR BAY    NEAR OCEAN
##         9136       6551        5       2290      2658
```

Tra queste 5 categorie si distingue indubbiamente “ISLAND”, presente in solo 5 distretti. Potrebbe essere sensato valutare di non considerare tale categoria, ma ciò si inserisce in un discorso molto più ampio, che comprende anche tutte le altre categorie. Nello specifico, dovremo considerare **ocean_proximity** non a livello di singola variabile, ma in relazione ai valori assunti da **Y** a seconda delle varie categorie. Tutto questo, ovviamente, lo potremo fare solo dopo aver confermato che effettivamente esista una qualche relazione tra la posizione geografica del distretto e il prezzo mediano delle case ad esso associato. Se non vi fosse alcuna relazione, infatti, potremmo decidere fin da subito di non includere le variabili geografiche nel nostro modello.

Il primo passo è dunque stabilire se esiste una qualche relazione tra la posizione geografica del distretto e la variabile risposta:



Appurato che una relazione esiste (*il grafico sopra suggerisce una relazione positiva tra Y e la vicinanza all’oceano*), si può procedere analizzando come si distribuisce **Y** nei vari livelli di **ocean_proximity**



Questo boxplot sembrerebbe confermare quanto intuito nel grafico precedente, ossia pare esserci una relazione positiva tra **Y** e la vicinanza all’oceano, e ciò è evidente soprattutto considerando le modalità “INLAND” e “ISLAND”. Nel primo caso il prezzo mediano delle abitazioni dell’entroterra appare in media molto più basso rispetto alle altre categorie; nelle isole invece vale il contrario, ma teniamo sempre presente questa categoria presenta una numerosità davvero minima. Per quanto riguarda invece le categorie “<1H OCEAN”, “NEAR BAY” e “NEAR OCEAN” la differenza in media non sembra così rilevante, e ciò potrebbe avere conseguenze a livello di nullità dei coefficienti del modello.

Tuttavia, **trarre conclusioni da un boxplot è un approccio troppo ingenuo**: è necessario approfondire la questione. A tal proposito, definiamo per $i = 1, \dots, n$ le seguenti variabili indicatrici:

$$x_{i2} = \begin{cases} 1 & \text{se l'unità } i \text{ ha ocean_proximity = "INLAND"} \\ 0 & \text{altrimenti} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{se l'unità } i \text{ ha ocean_proximity = "ISLAND"} \\ 0 & \text{altrimenti} \end{cases}$$

$$x_{i4} = \begin{cases} 1 & \text{se l'unità } i \text{ ha ocean_proximity = "NEAR BAY"} \\ 0 & \text{altrimenti} \end{cases}$$

$$x_{i5} = \begin{cases} 1 & \text{se l'unità } i \text{ ha ocean_proximity = "NEAR OCEAN"} \\ 0 & \text{altrimenti} \end{cases}$$

e consideriamo il modello $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i$. L'interpretazione dei parametri sarà:

- β_1 è la media di Y in un distretto appartenente alla classe “<1H OCEAN”;
- β_2 è la differenza tra la media di Y in un distretto appartenente a “INLAND” e la media di Y in un distretto appartenente a “<1H OCEAN”;
- β_3 è la differenza tra la media di Y in un distretto appartenente a “ISLAND” e la media di Y in un distretto appartenente a “<1H OCEAN”;
- β_4 è la differenza tra la media di Y in un distretto appartenente a “NEAR BAY” e la media di Y in un distretto appartenente a “<1H OCEAN”;
- β_5 è la differenza tra la media di Y in un distretto appartenente a “NEAR OCEAN” e la media di Y in un distretto appartenente a “<1H OCEAN”;

```
data$ocean_proximity = factor(data$ocean_proximity)
fit_ocean<- lm(Y~ocean_proximity, data=data)
summary(fit_ocean)
```

```
##
## Call:
## lm(formula = Y ~ ocean_proximity, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -236712  -66247  -21005   42273   375196 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 240084     1054 227.804 < 2e-16 ***
## ocean_proximityINLAND -115279     1631 -70.686 < 2e-16 ***
## ocean_proximityISLAND  140356     45062   3.115  0.00184 ** 
## ocean_proximityNEAR BAY 19128      2354   8.125 4.71e-16 ***
## ocean_proximityNEAR OCEAN 9350      2220   4.212 2.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 100700 on 20635 degrees of freedom
## Multiple R-squared:  0.2381, Adjusted R-squared:  0.238 
## F-statistic: 1612 on 4 and 20635 DF,  p-value: < 2.2e-16
```

Il test di nullità sui singoli coefficienti rivela che essi sono tutti significativamente diversi da 0. Ha dunque senso supporre che vi sia una differenza in media significativa tra i 5 gruppi. Inoltre, facciamo presente che oltre al p-value è sempre bene valutare anche lo standard error dei coefficienti: se lo s.e. è molto elevato e non c'è una particolare ragione di includere la variabile, potrebbe aver senso escluderla dal modello. Tra i nostri coefficienti, risulta uno s.e. molto elevato nel gruppo "ISLAND". Ricordiamo anche che tale categoria è presente in sole 5 osservazioni del dataset. Per questi due motivi, si potrebbe decidere di non includere tale variabile nel modello. Allo stesso tempo, ricordiamo che il grande dubbio che ci ha condotti a questa analisi riguardava le categorie "<1H OCEAN", "NEAR BAY" e "NEAR OCEAN". Se da un lato il boxplot non sembrava evidenziare differenze in media significative, dall'altro il test appena effettuato suggerirebbe il contrario. Possiamo dunque effettuare un terzo test sul confronto tra i seguenti modelli annidati:

$$M_0 : Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

$$M_1 : Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i$$

Si tratta dunque di verificare l'ipotesi nulla $H_0 : \beta_4 = \beta_5 = 0$ che, in termini, è del tutto equivalente all'ipotesi che il modello ridotto (M_0) sia migliore del modello completo (M_1).

Per farlo è necessario codificare **ocean_proximity** in modo diverso. Senza andare a modificare il dataset originale, ho valutato fosse opportuno creare un dataframe apposito per quest'analisi.

```
Y<- data$Y
inland<- data$ocean_proximity=="INLAND"
island<- data$ocean_proximity=="ISLAND"
nearbay<- data$ocean_proximity=="NEAR BAY"
nearocean<- data$ocean_proximity=="NEAR OCEAN"
ocean_data<- data.frame(cbind(Y, inland, island, nearbay, nearocean))
```

NB: la categoria "<1H OCEAN" non è stata inclusa, in modo da interpretarla come intercetta del modello!

A questo punto è stato possibile procedere al confronto tra i modelli M_0 ed M_1 , tramite **anova**:

```
M1= lm(Y~.,ocean_data)
M0= lm(Y~inland+island,ocean_data)
anova(M0,M1,test='F')
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ inland + island
## Model 2: Y ~ inland + island + nearbay + nearocean
##   Res.Df      RSS Df Sum of Sq    F    Pr(>F)
## 1  20637 2.1013e+14
## 2  20635 2.0939e+14  2 7.3555e+11 36.243 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

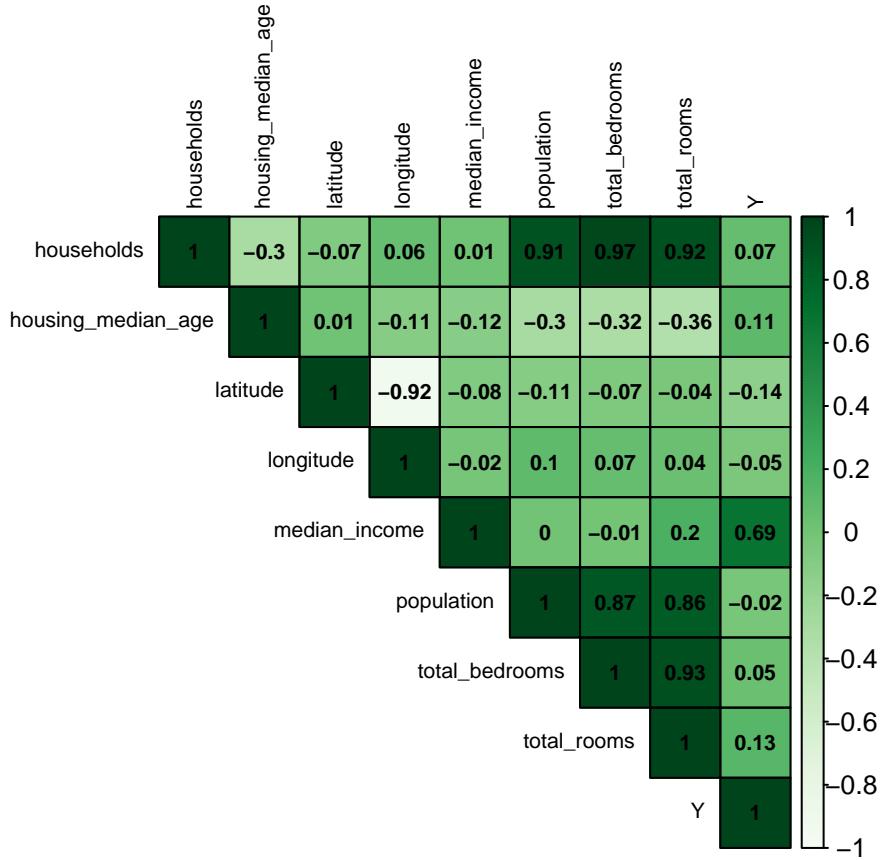
L'analisi della varianza rivela una forte evidenza contro l'ipotesi nulla, dunque tra i due modelli quello che risulta migliore è il modello completo M_1 . In altri termini, le differenze in media di **Y** tra le varie categorie sono tutte significative, anche se resta comunque una certa diffidenza verso la categoria "ISLAND".

NOTA: in generale, per poter applicare questi metodi, è necessario che a monte siano verificate alcune ipotesi. I dati a disposizione purtroppo non le rispettavano, quindi le conclusioni a cui siamo giunti devono essere considerate con cautela. Allo stesso tempo però, pur applicando trasformazioni varie la situazione non è migliorata: non avrei saputo cos'altro fare per poter valutare le differenze in media tra gruppi.

FEATURE SELECTION

Passiamo ora a una tra le fasi più complesse della costruzione del modello: la *selezione delle variabili*.

Capire che variabili includere nel modello è fondamentale, ma allo stesso tempo non esistono procedure standard per farlo. Spesso si sfruttano alcuni criteri di informazione (AIC, BIC, Cp); altre volte si confrontano direttamente i modelli. Tuttavia, alcune semplici accortezze in fase di analisi esplorativa possono semplificare notevolmente le cose nel seguito. Nello specifico, ho voluto risolvere a monte il problema della *multicollinearità*. Valutiamo dunque come sono correlate tra loro le variabili:



Concentriamoci innanzitutto sulla correlazione tra le sole variabili esplicative:

- **longitude** e **latitude** sono molto correlate, ma allo stesso tempo non avrebbe senso prenderle singolarmente;
- **population**, **households**, **total_bedrooms** e **total_rooms** sono variabili fortemente correlate, ed in realtà potevamo aspettarcelo;
- le altre variabili fortunatamente non presentano correlazioni rilevanti;

La correlazione tra **longitude** e **latitude** può essere facilmente evitata inserendo nel modello la variabile geografica **ocean_proximity**. È inoltre evidente che non si possono includere insieme le variabili **population**, **households**, **total_bedrooms** e **total_rooms**, ma come si può decidere cosa escludere?

La scelta non è così banale. Tutte le variabili sono legate alla dimensione della popolazione, ma escludere **population** non risolverebbe il problema, poiché **total_bedrooms** e **total_rooms** sarebbero comunque correlate a **households**. Escludere pure quest'ultima non aiuterebbe, perché il totale delle stanze sarebbe comunque legato al totale delle camere da letto (d'altronde sempre di stanze si tratta).

La soluzione sembrerebbe quindi selezionare una sola di queste 4 grandezze. In realtà si può fare di meglio!

FEATURE ENGINEERING

Il *feature engineering* è un processo di creazione, trasformazione e selezione delle variabili (features) presenti nei dati grezzi, e può influire notevolmente sulle prestazioni di un modello. Tale processo viene adottato in generale per fornire al modello informazioni più significative e, auspicabilmente, per migliorare la sua capacità previsiva. A partire dai dati, ho dunque costruito 3 nuove grandezze:

- 1) **rooms_per_household**: si tratta del rapporto tra il totale delle stanze e il numero di famiglie del distretto. Esso fornisce una misura delle stanze per famiglia nel distretto considerato;
- 2) **bedrooms_per_room**: si tratta del numero di camere da letto in rapporto al numero totale di stanze;
- 3) **pop_per_household**: si tratta della popolazione del distretto rapportata al numero di famiglie dello stesso. Esso fornisce una misura della densità di popolazione del distretto;

```
data$rooms_per_household<- data$total_rooms / data$households  
data$bedrooms_per_room<- data$total_bedrooms / data$total_rooms  
data$pop_per_household<- data$population / data$households
```

LE NUOVE VARIABILI SONO PIÙ SIGNIFICATIVE PER Y ?

Confronto tra **total_rooms** e **rooms_per_household**

```
cor(data$total_rooms, data$Y)
```

```
## [1] 0.1341531
```

```
cor(data$rooms_per_household, data$Y)
```

```
## [1] 0.1519483
```

La nuova variabile, anche se di poco, è più correlata alla variabile risposta!

Confronto tra **total_bedrooms** e **bedrooms_per_room**

```
cor(data$total_bedrooms, data$Y)
```

```
## [1] 0.04945686
```

```
cor(data$bedrooms_per_room, data$Y)
```

```
## [1] -0.2333029
```

La nuova variabile è decisamente più correlata alla variabile risposta. Inoltre, notiamo come il segno della correlazione sia diventato negativo, ossia al diminuire del rapporto camere/stanze il valore della casa aumenta. Ciò potrebbe avere senso, in quanto case dal valore elevato è probabile che abbiano stanze “accessorie” (*studi, bagni privati, mansarde,...*) che, aumentando il denominatore, riducono tale rapporto.

LE NUOVE VARIABILI SONO CORRELATE TRA LORO ?

```
cor(data$rooms_per_household, data$pop_per_household)
```

```
## [1] -0.004852295
```

```
cor(data$bedrooms_per_room, data$pop_per_household)
```

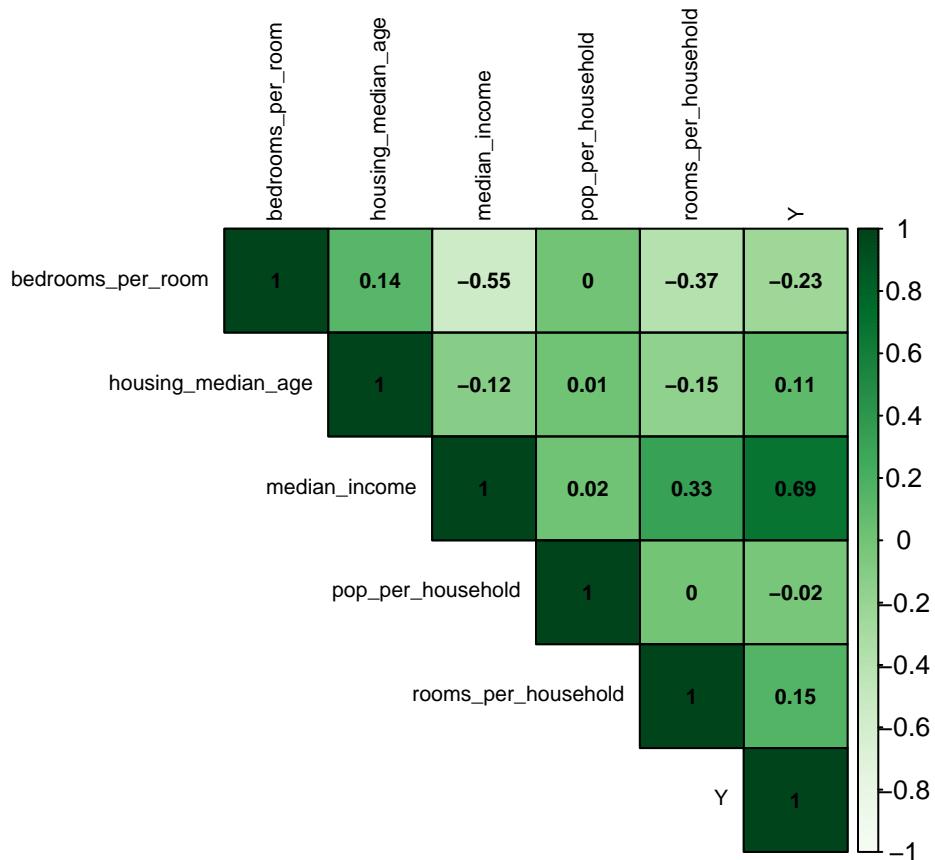
```
## [1] 0.002600952
```

```
cor(data$rooms_per_household, data$bedrooms_per_room)
```

```
## [1] -0.3703083
```

rooms_per_household e **bedrooms_per_room** sono pressochè incorrelate con **pop_per_household**, mentre in precedenza sia **total_rooms** sia **total_bedrooms** erano molto correlate a **population**. Inoltre, se prima **total_rooms** e **total_bedrooms** avevano una correlazione di circa 0.93, ora la correlazione tra **rooms_per_household** e **bedrooms_per_household** risulta soltanto -0.37.

Grazie a queste procedure preliminari abbiamo eliminato sul nascere i problemi di multicollinearità, e la situazione è nettamente migliorata. La matrice di correlazione tra le esplicative è diventata infatti:



MODELLO

```
# dividiamo test set e training set
set.seed(1)
sample<- sample(c(TRUE,FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
original_data<- data
test<- data[!sample,]
data<- data[sample,]
```

Modello con tutte le esplicative originali

```
fit_originale<- lm(Y~ longitude+latitude+housing_median_age+total_rooms+
                     total_bedrooms+households+population+median_income+
                     ocean_proximity, data=data)
require(MASS)
```

```
## Caricamento del pacchetto richiesto: MASS
```

```
fit_originale<- stepAIC(fit_originale, direction="both")
```

```
## Start: AIC=366695.3
## Y ~ longitude + latitude + housing_median_age + total_rooms +
##      total_bedrooms + households + population + median_income +
##      ocean_proximity
##
##          Df  Sum of Sq      RSS      AIC
## <none>                 7.8501e+13 366695
## - total_rooms            1 1.3818e+11 7.8639e+13 366722
## - households             1 4.9631e+11 7.8997e+13 366797
## - total_bedrooms          1 5.4196e+11 7.9043e+13 366807
## - housing_median_age     1 2.1696e+12 8.0670e+13 367142
## - latitude                1 2.2288e+12 8.0729e+13 367154
## - ocean_proximity         4 2.2584e+12 8.0759e+13 367154
## - longitude               1 2.4361e+12 8.0937e+13 367196
## - population              1 5.0327e+12 8.3533e+13 367716
## - median_income            1 5.1644e+13 1.3014e+14 375011
```

```
summary(fit_originale)
```

```
##
## Call:
## lm(formula = Y ~ longitude + latitude + housing_median_age +
##      total_rooms + total_bedrooms + households + population +
##      median_income + ocean_proximity, data = data)
##
## Residuals:
##    Min      1Q  Median      3Q      Max
## -550004 -43030 -10537  28926  759903
```

```

## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -2.187e+06  9.860e+04 -22.185 < 2e-16 ***
## longitude            -2.583e+04  1.144e+03 -22.587 < 2e-16 ***
## latitude             -2.442e+04  1.130e+03 -21.605 < 2e-16 ***
## housing_median_age    1.050e+03  4.928e+01  21.316 < 2e-16 ***
## total_rooms           -4.626e+00  8.599e-01 -5.379 7.57e-08 ***
## total_bedrooms        6.981e+01  6.553e+00  10.654 < 2e-16 ***
## households           7.480e+01  7.337e+00  10.195 < 2e-16 ***
## population            -3.778e+01  1.164e+00 -32.465 < 2e-16 ***
## median_income         3.867e+04  3.718e+02 103.998 < 2e-16 ***
## ocean_proximityINLAND -4.119e+04  1.967e+03 -20.939 < 2e-16 ***
## ocean_proximityISLAND  1.405e+05  3.459e+04  4.062 4.89e-05 ***
## ocean_proximityNEAR BAY -4.526e+03  2.142e+03 -2.113  0.03458 *  
## ocean_proximityNEAR OCEAN 4.807e+03  1.753e+03  2.742  0.00612 ** 
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 69100 on 16440 degrees of freedom
## Multiple R-squared:  0.6427, Adjusted R-squared:  0.6424 
## F-statistic:  2464 on 12 and 16440 DF,  p-value: < 2.2e-16

```

Questo modello sicuramente non è adatto, è stato inserito solamente come punto di partenza.

Osserviamo che:

- tutti i coefficienti sono significativi ad un livello di significatività $\alpha = 0.05$;
- la procedura *stepAIC()* ha mantenuto tutte le variabili, senza scartarne alcuna;
- l' R^2_{adj} è risultato 0.6424;
- come analizzato in precedenza, c'è forte presenza di multicollinearità;

Modello con le nuove variabili (e ocean_proximity)

```

fit_newvars<- lm(Y~ housing_median_age+median_income+rooms_per_household+
                    bedrooms_per_room+pop_per_household+
                    ocean_proximity, data=data)
summary(fit_newvars)

## 
## Call:
## lm(formula = Y ~ housing_median_age + median_income + rooms_per_household +
##     bedrooms_per_room + pop_per_household + ocean_proximity,
##     data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max  
## -739995 -44924 -11742  29823  472996 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -11513.61     4371.94  -2.634  0.00846 ** 
## 
```

```

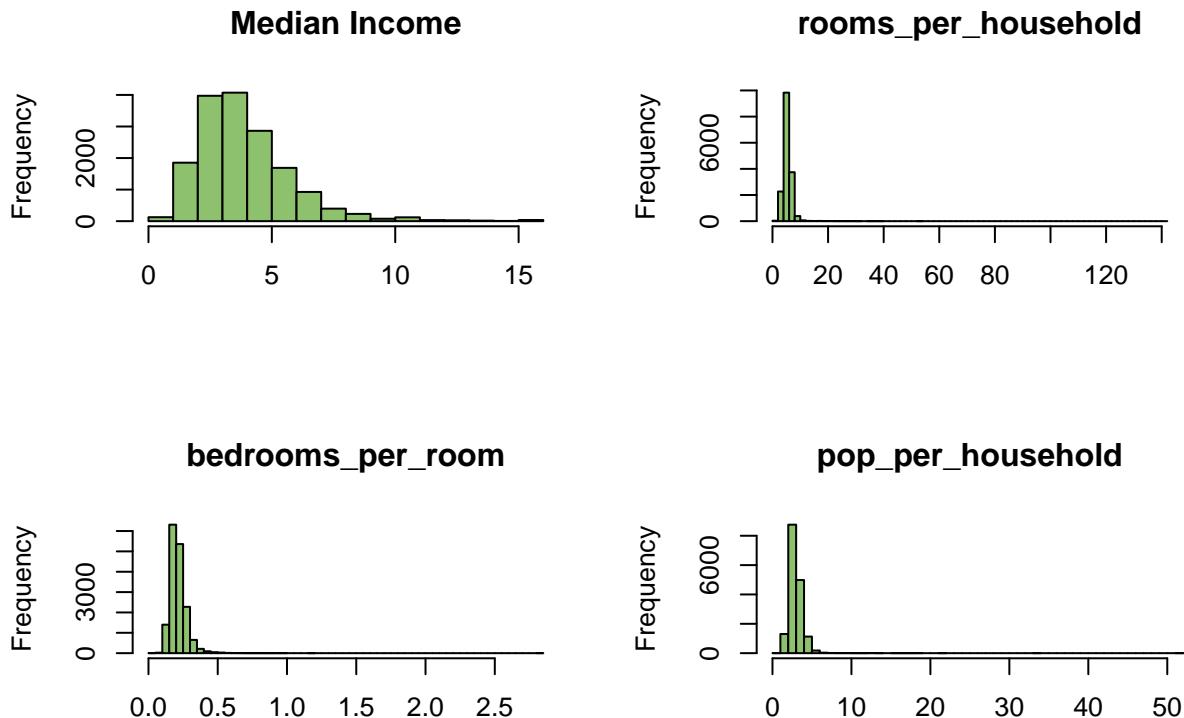
## housing_median_age          947.02      48.58 19.493 < 2e-16 ***
## median_income              41135.38     396.26 103.809 < 2e-16 ***
## rooms_per_household        1817.93      277.00  6.563 5.44e-11 ***
## bedrooms_per_room          191437.25    11238.76 17.034 < 2e-16 ***
## pop_per_household          -331.87      48.88 -6.790 1.16e-11 ***
## ocean_proximityINLAND     -67650.77     1468.43 -46.070 < 2e-16 ***
## ocean_proximityISLAND      161789.96    36396.04  4.445 8.84e-06 ***
## ocean_proximityNEAR BAY    13361.02      1943.55  6.875 6.44e-12 ***
## ocean_proximityNEAR OCEAN  18093.21      1789.48 10.111 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72750 on 16443 degrees of freedom
## Multiple R-squared:  0.6038, Adjusted R-squared:  0.6036
## F-statistic:  2784 on 9 and 16443 DF,  p-value: < 2.2e-16

```

- tutti i coefficienti sono significativi;
- rimane sempre il dubbio se includere la categoria “ISLAND” (s.e. molto elevato);
- la procedura *stepAIC* (anche se non inserita nel report) mantiene tutte le variabili;
- l’ R^2_{adj} è peggiorato rispetto a prima, risultando 0.6036 ;
- il modello non presenta problemi di multicollinearità;

Modello con le variabili trasformate

Nel modello precedente alcune esplicative hanno una distribuzione molto asimmetrica:



Passiamo dunque ai logaritmi (avevo provato anche altre combinazioni, ma il log è stata la migliore):

```

data$log_median_income<- log(data$median_income)
data$log_rooms_per_household<- log(data$rooms_per_household)
data$log_bedrooms_per_room<- log(data$bedrooms_per_room)
data$log_pop_per_household<- log(data$pop_per_household)
data$log_Y<- log(data$Y)
test$log_median_income<- log(test$median_income)
test$log_rooms_per_household<- log(test$rooms_per_household)
test$log_bedrooms_per_room<- log(test$bedrooms_per_room)
test$log_pop_per_household<- log(test$pop_per_household)
test$log_Y<- log(test$Y)

fit_log<- lm(log_Y~housing_median_age + log_median_income+ log_rooms_per_household+
              log_bedrooms_per_room+log_pop_per_household+ocean_proximity, data=data)
summary(fit_log)

## 
## Call:
## lm(formula = log_Y ~ housing_median_age + log_median_income +
##     log_rooms_per_household + log_bedrooms_per_room + log_pop_per_household +
##     ocean_proximity, data = data)
##
## Residuals:
##      Min        1Q        Median        3Q        Max 
## -2.46506 -0.20282 -0.01339  0.19047  2.98196 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           11.8881896  0.0222076 535.320 < 2e-16 ***
## housing_median_age    0.0028368  0.0002217 12.798 < 2e-16 ***
## log_median_income     0.7729136  0.0087389  88.445 < 2e-16 ***
## log_rooms_per_household 0.0437331  0.0154077  2.838 0.004540 **  
## log_bedrooms_per_room   0.2358573  0.0191198 12.336 < 2e-16 ***
## log_pop_per_household  -0.3808776  0.0096914 -39.301 < 2e-16 ***
## ocean_proximityINLAND -0.4678336  0.0070015 -66.819 < 2e-16 ***
## ocean_proximityISLAND   0.5829687  0.1649695   3.534 0.000411 *** 
## ocean_proximityNEAR BAY -0.0128697  0.0089385  -1.440 0.149941  
## ocean_proximityNEAR OCEAN  0.0097334  0.0081754   1.191 0.233843 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3296 on 16443 degrees of freedom
## Multiple R-squared:  0.6664, Adjusted R-squared:  0.6662 
## F-statistic: 3650 on 9 and 16443 DF,  p-value: < 2.2e-16

```

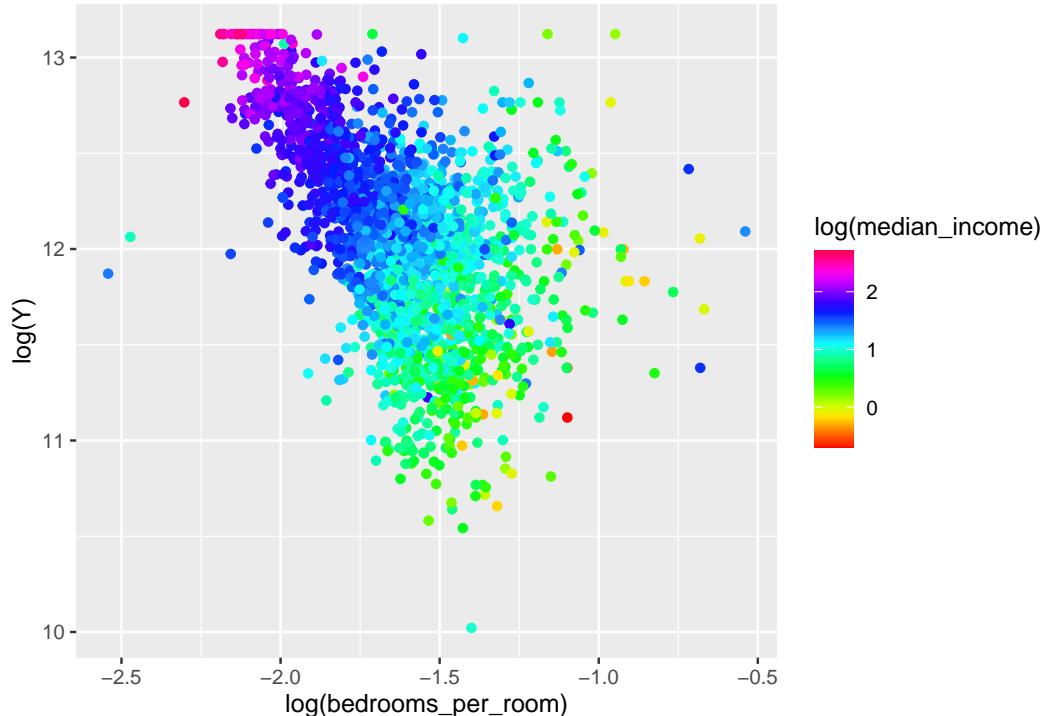
- “NEAR BAY” e “NEAR_OCEAN” non sono significativi;
- $\log_{\text{rooms}}_{\text{per}}_{\text{household}}$ è significativo usando un livello di significatività da 0.01 in poi;
- l’ R^2_{adj} è migliorato rispetto ad entrambi i modelli precedenti, risultando 0.6662;
- NB: avevo provato anche a centrare le esplicative nella media $[X'_i = X_i - \bar{X}_i]$, ma non è cambiato nulla
- avevo anche provato a standardizzare $[X'_i = \frac{X_i - E(X_i)}{\sqrt{V(X_i)}}]$, ma ancora niente!
- visto che alcuni coefficienti di ocean erano non significativi, avevo provato a includere longitudine e latitudine sottoforma di rapporto (per evitare collinearità), ma il modello peggiorava significativamente.

PARADOSSO DI SIMPSON

Si ha un paradosso di Simpson quando i dati mostrano un'associazione di un certo tipo tra due variabili se li si guarda condizionatamente ad una terza, ma mostrano l'associazione opposta se si ignora la terza variabile.

Il coefficiente associato a **log_bedrooms_per_room** è > 0 , dunque l'effetto di tale esplicativa sulla **Y** risulterebbe positivo. Tuttavia, avevamo analizzato in precedenza la correlazione negativa tra queste due variabili, interpretandola come sensata. Allora come mai il coefficiente è positivo?

Effettivamente, se si costruisce il modello partendo da **log_bedrooms_per_room** e si aggiungono man mano le altre esplicative, il coefficiente di **log_bedrooms_per_room** rimane sempre negativo. Nel momento in cui però si include anche la variabile **log_median_income**, ecco che il coefficiente cambia segno. Si tratta di un chiaro esempio di paradosso di Simpson, e possiamo analizzarlo meglio con il seguente grafico:



Commento: se consideriamo i dati a livello complessivo, la correlazione tra **log_bedrooms_per_room** e **log_Y** è negativa. Se ci poniamo sui particolari livelli di reddito, invece, la relazione tra le due variabili diventa positiva. Inoltre, possiamo apprezzare come per i livelli centrali del reddito questa correlazione sia molto marcata, mentre nei livelli di reddito più alti essa risulti meno evidente (è sensato che sia così!)

Modello solo quantitativo

Visto che ben 2 categorie di **ocean_proximity** non risultavano significative, e visto che la cartina mostrava una netta corrispondenza tra la **Y** e la vicinanza alla costa, ho arricchito il mio dataset con una nuova grandezza: **coast_distance**. Si tratta di una misura quantitativa della distanza di ogni distretto dalla costa più vicina. Nello specifico, è stata calcolata tramite la formula di Haversine, che date le coordinate di due punti na calcola la distanza (in km) come:

$$d = 2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right), \text{ in cui } \begin{cases} r : \text{raggio terrestre (6371Km)} \\ \phi_1, \phi_2 : \text{latitudini dei punti} \\ \lambda_1, \lambda_2 : \text{longitudini dei punti} \end{cases}$$

Anche se le coordinate sono dati pubblici, non ho dovuto effettuare tale calcolo. Ho sfruttato un dataset presente su Kaggle in cui i calcoli erano già svolti:

```
dati<- read.csv("data/california_distances.csv", header=T)
data$coast_distance<- dati$Distance_to_coast[sample]
test$coast_distance<- dati$Distance_to_coast[!sample]
original_data$coast_distance<- dati$Distance_to_coast
```

Proviamo a costruire un modello simile al precedente, ma con la radice della distanza dalla costa (e radice di median_income):

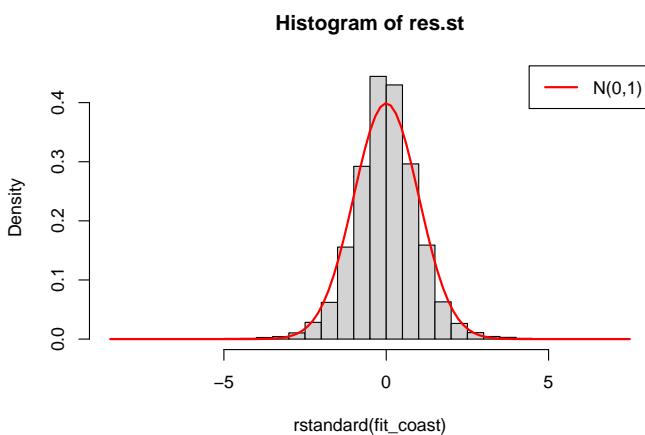
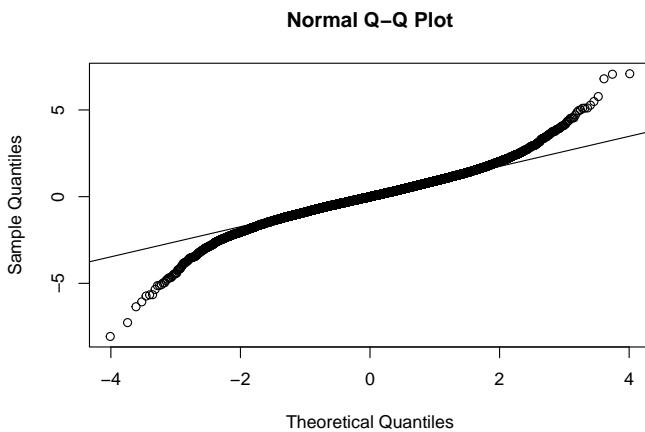
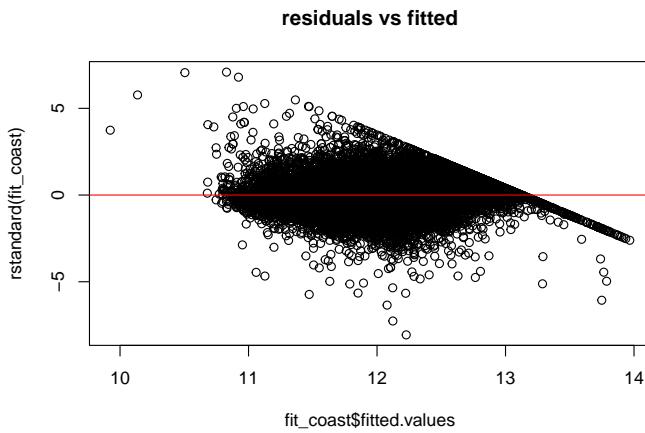
```
fit_coast<- lm(log(Y)~housing_median_age + sqrt(median_income)+ log_rooms_per_household+
log_bedrooms_per_room+log_pop_per_household+sqrt(coast_distance), data=data)
summary(fit_coast)

##
## Call:
## lm(formula = log(Y) ~ housing_median_age + sqrt(median_income) +
##     log_rooms_per_household + log_bedrooms_per_room + log_pop_per_household +
##     sqrt(coast_distance), data = data)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -2.61150 -0.18822 -0.00135  0.19142  2.29380
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.154e+01  2.141e-02 539.110 < 2e-16 ***
## housing_median_age        2.153e-03  2.141e-04 10.058 < 2e-16 ***
## sqrt(median_income)       8.381e-01  8.965e-03 93.483 < 2e-16 ***
## log_rooms_per_household   6.759e-02  1.532e-02  4.412 1.03e-05 ***
## log_bedrooms_per_room     3.269e-01  1.863e-02 17.550 < 2e-16 ***
## log_pop_per_household    -3.411e-01  9.371e-03 -36.401 < 2e-16 ***
## sqrt(coast_distance)     -2.095e-03  2.935e-05 -71.372 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.324 on 16446 degrees of freedom
## Multiple R-squared:  0.6776, Adjusted R-squared:  0.6775
## F-statistic:  5760 on 6 and 16446 DF,  p-value: < 2.2e-16
```

Tutti i coefficienti risultano significativi e l' R^2_{adj} è aumentato a 0.6775

Analisi dei residui

L'analisi dei residui dei modelli precedenti non è risultata soddisfacente, oltre che poco chiara visto l'elevato numero di dati. Propongo ora l'analisi dei residui di quest'ultimo modello (il migliore in termini di R^2_{adj}):

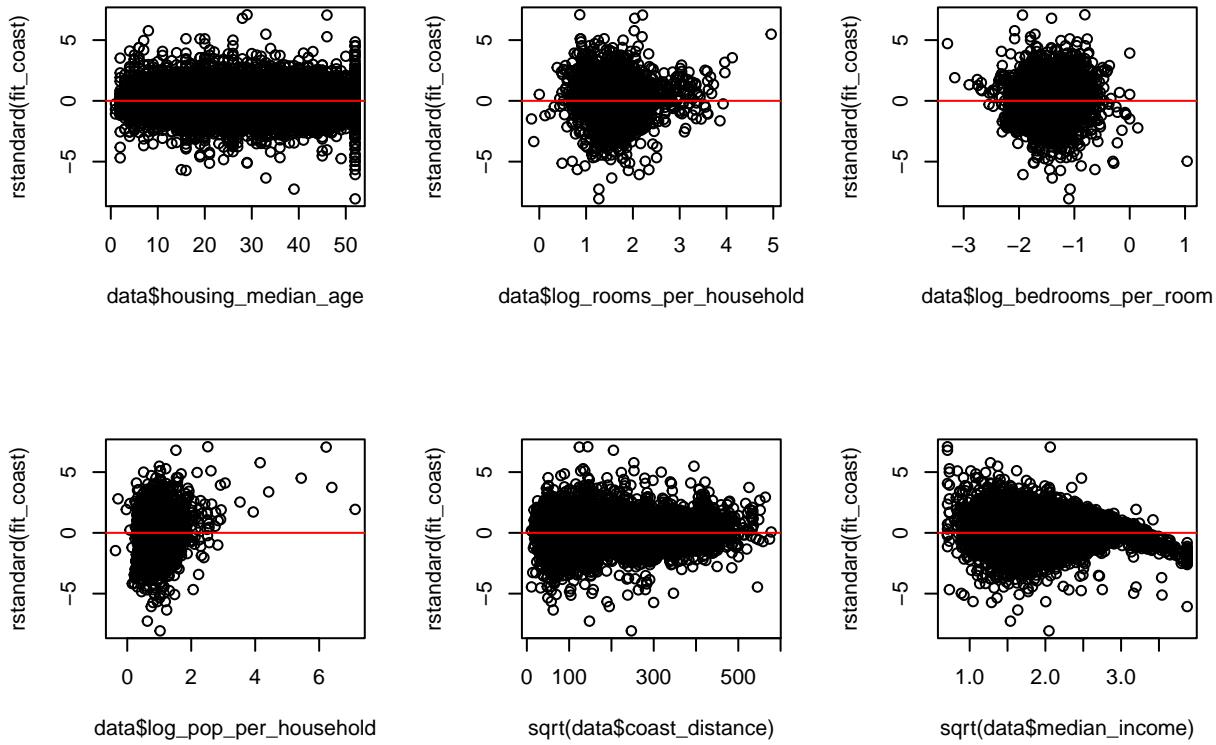


I residui del modello non sono molto soddisfacenti. D'altronde il modello spiega meno del 70% della varianza. Particolamente interessante è la "linea diagonale" che si nota nel grafico "residuals vs fitted". Non ho trovato una spiegazione sensata, ma forse può essere una conseguenza del threshold effect sulla Y (discusso nella parte iniziale). Inoltre, c'è da dire che l'elevata numerosità dei dati non rende di facile interpretazione i test grafici. Ho pensato che i risultati potessero essere influenzati molto anche la presenza di outliers, dunque ho provato a costruire il modello senza considerare i valori estremi di ciascuna colonna.

Per farlo, ho implementato la seguente funzione:

```
detect_outliers<- function(v){
  quantiles= quantile(v, names=F)
  W= quantiles[4]-quantiles[2]
  min= quantiles[2]- 1.5*W
  max= quantiles[4]+ 1.5*W
  return(v< min | v > max)
}
```

Pur rimuovendo tutti i record contenenti gli outliers, il modello non presentava miglioramenti significativi. L'ultimo tentativo per cercare di migliorare il modello è stato analizzare i residui rispetto a ciascuna esplicativa, in modo da verificare se fossero presenti alcuni andamenti sistematici rispetto alle singole variabili.



CROSS VALIDATION: TRAIN-TEST SET

Visto che i residui non sono di facile interpretazione, valutiamo la bontà del modello in termini quantitativi. Nello specifico, confrontiamo come i due modelli migliori performano sul test set (di dimensione T), valutando:

$$\text{MAE} = \frac{1}{T} \sum_{i=1}^T |y_i - \hat{y}_i| \quad ; \quad \text{RMSE} = \sqrt{\frac{1}{T} \sum_{i=1}^T (y_i - \hat{y}_i)^2}$$

```
previsioni_log<- predict(fit_log, test)
previsioni_coast<- predict(fit_coast, test)
MAE_log<- mean(abs(previsioni_log-log(test$Y)))
MAE_coast<- mean(abs(previsioni_coast-log(test$Y)))
RMSE_log<- sqrt(mean((previsioni_log-log(test$Y))^2))
RMSE_coast<- sqrt(mean((previsioni_coast-log(test$Y))^2))
```

```
## MAE del modello fit_log: 0.2418378
## MAE del modello fit_coast: 0.2359067
## RMSE del modello fit_log: 0.3189007
## RMSE del modello fit_coast: 0.3097795
```

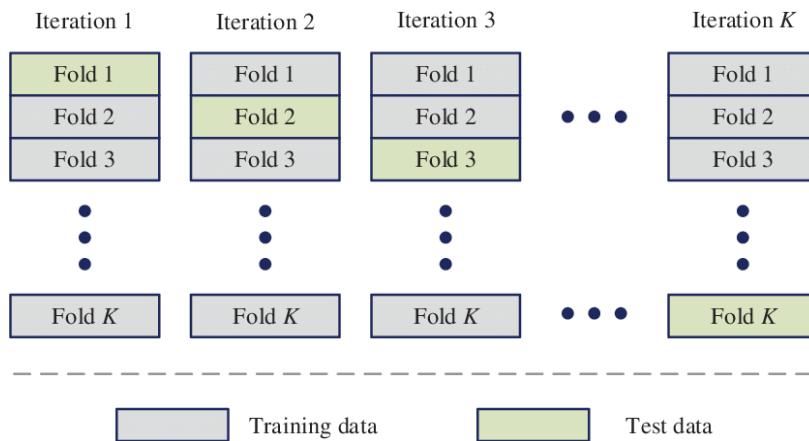
Il modello *fit_coast* risulta migliore di *fit_log* per R^2_{adj} , RMSE e MAE. Teniamo però presente che il risultato potrebbe dipendere anche dalla scelta del training set e del test set. Allora come possiamo essere sicuri che il risultato sia valido e che non sia stata fortuna? Esiste un approccio più stabile: la K-fold validation!

K-FOLD VALIDATION

La K-fold validation, a differenza della semplice divisione train-test set, utilizza più suddivisioni dei dati e calcola la media delle prestazioni per ottenere una valutazione del modello più precisa e robusta.

L'idea alla base è la seguente:

- 1) Si divide il campione in K parti (approssimativamente) uguali
- 2) Si esegue la procedura di train-test set usando ciascuna delle K parti come insieme di test (e le restanti $K-1$ come insieme di train)
- 3) Al termine, si avranno K stime dell'errore di previsione: la media di queste sarà la stima complessiva



Importiamo la librerie necessarie per automatizzare la K-fold validation:

```
library(lattice)
library(caret)
```

Impostiamo un *seed* in modo da poter (eventualmente) riprodurre la cross validation con gli stessi valori:

```
set.seed(1)
```

A questo punto si può procedere con la cross validation dei due modelli (con $K = 10$ per esempio):

```
train_control <- trainControl(method = "cv",
                                number = 10)

model_log <- train(log(Y)~housing_median_age + log(median_income)+ log(rooms_per_household)+
                    log(bedrooms_per_room)+log(pop_per_household)+sqrt(coast_distance),
                    data = original_data,
                    method = "lm",
                    trControl = train_control)

model_coast <- train(log(Y)~housing_median_age + sqrt(median_income)+ log(rooms_per_household)+
                     log(bedrooms_per_room)+log(pop_per_household)+ocean_proximity,
                     data = original_data,
                     method = "lm",
                     trControl = train_control)
```

I risultati sono stati i seguenti:

```
print(model_log)

## Linear Regression
##
## 20640 samples
##      6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 18575, 18576, 18576, 18576, 18576, 18576, ...
## Resampling results:
##
##    RMSE      Rsquared     MAE
##    0.3275375  0.6689782  0.2463837
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

print(model_coast)

## Linear Regression
##
## 20640 samples
##      6 predictor
```

```

## 
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 18577, 18576, 18576, 18574, 18575, 18577, ...
## Resampling results:
##
##    RMSE      Rsquared     MAE
##    0.3212698  0.6816054  0.241057
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

```

Anche utilizzando la K-fold validation il modello `fit_coast` è risultato migliore di `fit_log` sotto tutte e 3 le metriche considerate. Teniamo presente che però il modello è stato costruito su scala logaritmica di Y, quindi non è possibile interpretare il MAE e l'RMSE direttamente in dollari.

EXPLAINABILITY

Spesso al business che ci commissiona un progetto non interessa uno score tecnico come l' R^2_{adj} . Bisogna spiegare perché il modello è affidabile e perchè porta a determinate previsioni, e spesso il modo più efficace è farlo quantificarlo in termini di denaro.

Una buona misura potrebbe dunque essere fornita da MAE e RMSE su scala originale.

NB: ovviamente non basta fare l'esponenziale di MAE e RMSE del modello su scala logaritmica!

Consideriamo dunque il modello `model_coast` (ottenuto tramite cross validation) e, a titolo di esempio, consideriamo il MAE:

```

previsioni_coast_originalscale<- exp(predict(model_coast,test))
MAE_coast_originalscale<- mean(abs(previsioni_coast_originalscale- test$Y))

```

```
## MAE su scala originale: 47535.61 $
```

Commento: se utilizziamo il modello per stimare il valore mediano delle abitazioni di un distretto della California, dobbiamo mettere in conto che la valutazione potrebbe essere errata mediamente di 47535.61 dollari. Un errore sicuramente alto, ma teniamo presente che:

- il modello presentava un $R^2_{adj} = 0.6816$;
- la Y è una statistica complessiva (la mediana) del distretto. Probabilmente, un modello a livello di singola abitazione (con dati anche sulle caratteristiche della casa), avrebbe un'accuratezza maggiore;
- i residui non risultavano soddisfacenti;
- in generale il modello lineare ha il vantaggio di essere facilmente interpretabile (*Glass Box Models*), ma spesso ha un'accuratezza più bassa di modelli di regressione più sofisticati e *Black Box*. Per questo particolare dataset, modelli di machine learning come Decision Tree e Random Forest permettono di migliorare notevolmente l' R^2_{adj} .