PROGETTO STATISTICA

GRUPPO “IL PLETTRO MAGICO”

Nell’ambito del corso di “Statistica per Big Data” ci è stata proposta la presentazione di un progetto nel quale ci veniva richiesto di analizzare un dataset, gentilmente messoci a disposizione dal professor Francesco Dotto nei materiali didattici, e di ricavare da esso informazioni che potessero raccontare un qualche aspetto della popolazione di interesse.
Nonostante la difficile e per certi versi (volutamente) vaga consegna, abbiamo fatto del nostro meglio, e questo paper che ora seguirà è il risultato del nostro lavoro, frutto di faticose ore di studio del dataset e di ricerca in rete, per riuscire ad interpretare al meglio i dati e riuscire a raccontare, più o meno bene, qualcosa tramite essi.
Vorremmo inizialmente partire da una veloce analisi dei dati “grezzi” presenti nel dataset. In particolare, ci siamo trovati di fronte ad un totale di 89 variabili differenti (riguardanti i più disparati ambiti, dalla condizione di salute alla grandezza del nucleo familiare, dall’income medio del nucleo familiare alla quantità di sigarette fumate giornalmente), riguardanti un totale di 40.944 unità.
Importante sottolineare che non tutte le variabili sono disponibili per tutte le unità statistiche.
Da qui inizia il nostro percorso di analisi che, naturalmente, si è concentrato dapprima sullo studio delle variabili disponibili e poi, in un secondo momento, sullo studio delle variabili che vennero raccolte durante la somministrazione, nel 2017, del questionario alle nostre unità statistiche (che, per essere più precisi, erano tutti cittadini sudafricani residenti in Sud Africa).
Abbiamo pensato di approfondire, in particolare, la durata del matrimonio e della convivenza, scegliendo un insieme di variabili che risultassero attinenti ai nostri obiettivi.
Non avendo trovato però nel nostro dataset tutte le variabili di nostro interesse (variabili che, successivamente, andrò ad elencare), abbiamo avuto la necessità di richiedere al professore di metterci a disposizione queste ulteriori variabili che pure erano state, durante la raccolta dei dati, richieste alle unità statistiche.
In particolare, abbiamo dovuto richiedere le variabili B4.2 mar, B5.1 mary_m e B5.2 mary_l.).
Qui inizia il lavoro vero e proprio, che si articola in più sezioni.
La prima sezione riguarda la pulizia dei dati, da noi effettuata nel primo script denominato “Regressione logistica e pulizia dei dati”. Seppure molto complessa, questa operazione si è resa necessaria per una corretta analisi statistica e per mantenere coerenza e chiarezza nella struttura del progetto, lavorando con dataset contenenti solo le informazioni e le variabili di nostro interesse.
Abbiamo, dunque, creato un nuovo dataset, denominato “dataset_pulito”, comprendente 14 variabili, attraverso la funzione: *data.frame()*.
Successivamente abbiamo generato una nuova variabile “anni_convivenza”, nata dall’unione delle due variabili mary_m e mary_l che rappresentavano rispettivamente il numero di anni di matrimonio e il numero di anni di convivenza. Dal momento che queste due variabili sono complementari, ad ogni unità avrà, per una delle due variabili un valore NA, mentre per l'altra un valore numerico. Con la funzione *coalesce* abbiamo quindi "unito" le due colonne nella colonna anni_convivenza, associando a quest'ultima il valore non nullo tra queste due variabili.

Tuttavia nell’approfondire meglio il dataset pulito, ci siamo accorti di come la variabile “best_edu”, nonostante sia una variabile numerica discreta, non rappresentasse

effettivamente gli anni di studio di ogni unità statistica, bensì il loro grado di istruzione. Dopo un'approfondita ricerca su internet siamo riusciti, quindi, crearci una tabella di conversione tra gradi di istruzione e anni di studio in Sud Africa ed abbiamo, direttamente nel dataset_pulito, creato una colonna aggiuntiva, denominata anni_convivenza, in cui abbiamo convertito tutti i valori (da 1 a 35) di best_edu con i rispettivi anni di studio (con il comando ripetuto *ifelse* per ogni unità). Per visionare questa tabella di conversione, rimandiamo allo script, dove è presente la suddetta.
Abbiamo fatto queste operazioni perché lo scopo di questa nostra analisi è di riuscire a stimare la durata del matrimonio/convivenza delle unità statistiche, a partire da fattori quali gli anni di studio, l'income, la grandezza del nucleo familiare, l'età, l'etnia, la religione e il luogo di residenza.
Avevamo intenzione di aggiungere anche la variabile "Monthly_hours", ovvero le ore di lavoro in un mese della singola unità statistica, ma da una precedente analisi (da noi svolta) ci siamo accorti di come questa variabile non fosse statisticamente significativa in nessun caso. Abbiamo quindi reputato opportuno non considerarla nella presente analisi. Inoltre abbiamo deciso di non considerarla anche perché, oltre a non essere rilevante per il nostro studio, andava a ridurre notevolmente (più del 50%) il campione statistico sul quale ci è possibile lavorare. Questa significativa riduzione viene riportata sul nostro script nelle due somme alle righe #127 e #132, che ci evidenziano come il campione, senza tener conto di monthly_hours, sia di 7817 unità, mentre nel caso tenessimo in considerazione quest'ultima esso si ridurrebbe a 3093 unità.
Andando avanti, abbiamo preso poi in esame le variabili cardine del nostro progetto, fondamenta di tutto ciò che andremo ad approfondire nella seconda parte. In particolare, esse sono:

1. "married": variabile binomiale che ci comunica se l'unità statistica è sposata/convivente (1) oppure non lo è (0).
2. "mar": variabile che può assumere 3 valori differenti:
   "mar == 1": l'individuo risulta formalmente sposato,
   "mar == 2": l'individuo risulta convivente
   "mar == 3": l'individuo risulta sposato ma allo stesso tempo non convivente con il proprio partner.

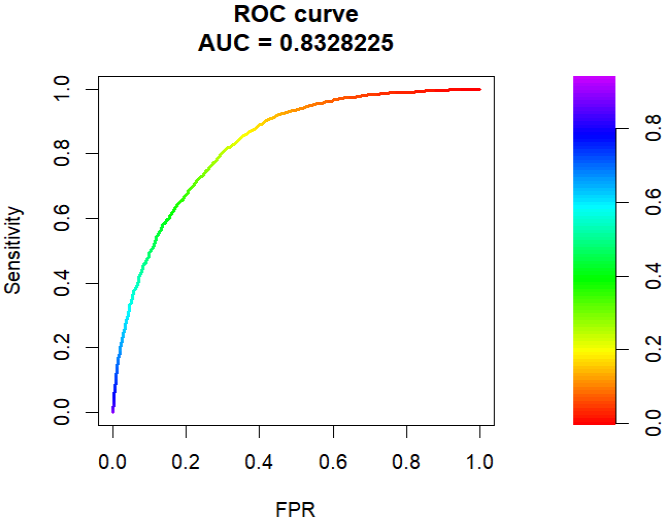
Studiando queste ultime, ci siamo soffermati su alcune incongruenze individuate nella loro intersezione, e ci siamo accorti come alcune unità statistiche non abbiano correttamente compilato il questionario proposto, andando di fatto ad alterare la coerenza del proprio status matrimoniale/di convivenza.
Queste incongruenze sono molteplici: abbiamo riscontrato sia individui che, nonostante abbiano un valore married == 1, non presentano alcun valore mar (abbiamo ricavato l'esatto numero di questo sottogruppo, grazie al comando *sum* a riga #145, che ci ha restituito ben 1665 unità), ma abbiamo anche riscontrato individui che nella variabile married sono classificati come non sposati e non conviventi (married == 0), ma hanno compilato anche la sezione del questionario che non gli spettava, ovvero la variabile mar, da compilare solo se appartieni a coloro che risultato essere married==1. Ben 243 unità in totale rientrano in questo gruppo, ed in particolare 121 di essi risultano essere mar == 3, mentre i restanti 122 risultano essere mar == 2.
Annotate queste piccole incongruenze iniziali, ci siamo immersi nel processo di pulizia dei dati vero e proprio. L'idea che abbiamo seguito è la seguente: tra le variabili non numeriche di nostro interesse (quelle nel dataset_pulito), l'unica, che come si vedrà in seguito può essere considerata una binomiale, che ci interessa andare a stimare partendo dalle restanti

variabili nel dataset (pulito), sapendo che le unità statistiche sono married == 1 (sposati o conviventi), riguarda la variabile mar, indicante l'essere sposato o semplicemente convivente.
Studiandoci la variabile mar più nel dettaglio, siamo giunti alla conclusione per la quale il nostro interesse si deve focalizzare solamente sui primi due valori della variabile, ovvero mar == 1 e mar == 2, ed in tal modo possiamo andiamo a trattare la variabile mar come una binomiale. Infatti noi andremo a scartare la terza informazione della variabile, che abbiamo ritenuto superflua per i nostri fini, in quanto se le unità non vivono insieme (mar==3) sarebbe inutile prenderle in considerazione nella stima della durata di una convivenza/matrimonio. Questa condizione è soddisfatta nelle unità che hanno married == 1 e mar == 3. Inoltre, la frequenza con cui tale occasione si verifica è talmente bassa da risultare insignificante (più precisamente le unità che sono contemporaneamente sposate ma non vivono insieme ammontano ad un totale di 20, ovvero solamente lo 0,048% su un campione iniziale di 40k osservazioni (si rimanda al comando a riga 125*)). Detto questo abbiamo quindi assunto che la nostra variabile mar, qualora married==1, possa essere solo mar==1 e mar==2.
Dopo aver chiarito perché "mar" può essere considerata una variabile binomiale, abbiamo notato la presenza di numerose unità statistiche in cui mar==NA (1665) e abbiamo deciso, quindi, di effettuare una regressione logistica, che ci andasse a stimare, ove possibile, il valore della variabile mar. Per fare questo abbiamo avuto la necessità di generare un nuovo dataset, che comprendesse le unità statistiche aventi tutti valori non nulli nelle variabili numeriche di nostro interesse per lo svolgimento della regressione logistica, ovvero: ppincome, hhsizer, age, anni_convivenza e anni_istruzione. Abbiamo anche ovviamente imposto che "mar" sia disponibile, sia diverso da 3 e che al contempo married sia pari ad 1.
Il nostro scopo è quindi ora quello di stimare, attraverso un campione completo di 6950 unità, i valori della variabile binomiale mar di tutte quelle unità che, nonostante presentino un profilo completo per le variabili indipendenti, la relativa variabile mar non assume alcun valore. Esse sono 867. Da notare che non è assolutamente detto che queste 867 variabili abbiano anche tutte quelle informazioni che ci serviranno per andare ad eseguire le regressioni multiple che seguiranno. La somma delle unità viene riportata a riga #179.

Regressione logistica

Prima di addentrarci nella mera esecuzione della nostra regressione logistica, ricordiamo brevemente che la regressione logistica è una tecnica statistica utilizzata per modellare la probabilità di un evento binario presupponendo una distribuzione binomiale per quanto riguarda la variabile risposta Y, che può assumere solamente due esiti (in questo caso mar==1 e mar ==2). I valori stimati dal modello logistico sono le probabilità che Y assuma l'uno o l'altro esito. Spetta a noi, inoltre, ricavare la soglia P ottimale che massimizza l'area sottostante alla curva ROC e che quindi rende il nostro modello di regressione il più accurato possibile.
Una volta trovata l'equazione della nostra regressione logistica (i cui coefficienti sono facilmente visionabili eseguendo la summary alla riga #188 del primo script), il nostro codice va a calcolare per ogni singola unità la la probabilità che "mar" sia 2 piuttosto che 1.
Abbiamo poi creato una colonna "mar_stimato", in cui andremo ad inserire il risultato della nostra regressione.
In particolare, ogni output va confrontato singolarmente con la soglia P ottimale, e nel caso esso risulti essere maggiore di quest'ultima, ad esso verrà assegnato un valore

mar_stimato==2, mentre in caso contrario ad esso verrà assegnato un valore mar_stimato==1.
Superfluo specificare che la soglia p ottimale è quella determinata soglia per cui i valori predetti risultano il più possibile simili ai valori reali nel campione analizzato, ed è quella soglia che non solo massimizza l'area sottostante la curva roc, ma massimizza anche la somma tra sensibilità e specificità.
Avendo studiato solamente la regressione logistica restituente una variabile dicotomica date delle variabili indipendenti di tipo numerico, noi siamo stati in grado di utilizzare come variabili indipendenti solamente ppincome, hhsizer, age, anni_convivenza e anni_istruzione.
Nello script sono comunque presenti, di tanto in tanto, dei commenti alle nostre operazioni, quindi rimandiamo alla visione dello stesso per una più completa analisi delle operazioni da noi effettuate.
Andiamo ora, dopo aver spiegato quello che abbiamo fatto, a commentare i risultati ottenuti.
Effettuando il comando *summary* (della nostra regressione logistica), possiamo subito notare come le variabili indipendenti da noi prese in considerazione siano tutte statisticamente significative (basta verificare che il p-value sia inferiore a 0.05).
Andando ora a commentare la curva ROC del nostro modello, ciò che a noi interessa è l'AUC (Area Under the Curve) ,che rappresenta la quanto buono siano le performance del nostro modello. Essa è 0.833, e il nostro modello risulta quindi essere un ottimo modello, che classifica correttamente le unità, per quel che riguarda la variabile mar, nell'83% dei casi.



Il grafico proposto mette in relazione "sensibilità" sulle ordinate e "1 - specificità" sulle ascisse, dove per sensibilità si intende quante volte il modello è stato in grado di azzeccare gli individui classificati come conviventi sul totale della popolazione che realmente convive, invece per specificità si intende la percentuale di sposati predetti dal modello sul totale di individui sposati. Di conseguenza "1-specificità" descrive i "False Negative", ovvero coloro che risultano essere conviventi ma che tuttavia il modello li ha classificati come sposati. I valori di sensibilità e specificità sono rispettivamente 0,7452 e 0,7472, e tutto ciò viene evidenziato dalla seguente matrice di confusione:

```
mar     1    2
    1 3870 1309
    2  451 1320
```

In particolare, il numero di unità che il modello è stato in grado di classificare correttamente (sia per ciò che riguarda il matrimonio che la convivenza) viene riportato sulla diagonale principale di questa matrice quadrata.
Inoltre siamo riusciti ad individuare la soglia p ottimale (p.opt), pari a 0.291, che ci consente di classificare nel migliore modo possibile le unità statistiche con la variabile mar mancante.
In altre parole, se il risultato della regressione logistica dovesse venire maggiore di p.opt allora sarebbe più probabile che l'individuo stia semplicemente convivendo rispetto a che sia

sposato, e verrà quindi associato, nella colonna mar_stimato, un valore pari a 2. Viceversa nel caso p.opt risulti essere minore.
Con la nostra regressione logistica abbiamo quindi integrato la variabile mar per 867 unità statistiche, alcune delle quali saranno poi prese da noi in considerazione nelle nostre regressioni multiple che seguiranno.
Riprendendo il discorso riguardante le incongruenze e i "problemi" che abbiamo riscontrato nel nostro dataset, ci siamo focalizzati ora su un'ulteriore variabile che avremmo voluto prendere in considerazione nella stima della durata del matrimonio o della convivenza, ovvero la variabile "em1".
Questa variabile presenta infatti due esiti differenti: 1 se l'individuo viene pagato regolarmente e 2 se invece l'individuo non viene pagato regolarmente, e se ci si stesse chiedendo, a questo punto, se si possa fare un ragionamento simile a quello precedentemente effettuato per la stima della variabile mar con la regressione logistica anche per stimare em1, qualora presenti valori nulli (dato che sembrerebbe essere una normale variabile dicotomica), la risposta è negativa. Abbiamo infatti approfondito la natura stessa della variabile em1 e ci siamo accorti in realtà che l'esito NA non risulta essere tale solo perché vi sono dei dati mancanti o la sezione è incompleta, ma la variabile potrebbe risultare em1 anche perché l'individuo non è un lavoratore dipendente, ed è, ad esempio, un imprenditore, un libero professionista, un disoccupato o semplicemente appartenente a tutte quelle professioni che non risultano essere subordinate. Dunque non avrebbe senso andare a stimare em1 per gli individui per cui em1==NA, perchè andremmo automaticamente a scartare la possibilità che il dato sia na a causa di un impiego lavorativo dell'unità statistica di tipo non subordinato.
Appurato che quindi non ha alcun senso andare ad effettuare una regressione logistica per la variabile em1, è importante notare che se noi prendessimo in considerazione nelle nostre future regressioni l'indicatore em1, andremmo a restringere la nostra analisi e le nostre conclusioni solo agli individui che risultano essere lavoratori subordinati, il che non è nostro interesse.
Nonostante inoltre potrebbe essere interessante conoscere la stabilità economica di una famiglia o di un individuo, per capire se esso si sente "tranquillo" o meno con il proprio stipendio, non è stata raccolta, e quindi non è a noi disponibile, una variabile abbastanza comprensiva. Nel caso di ulteriori dubbi sul fatto che la variabile em1 abbia ricevuto risposta solo dai lavoratori dipendenti, rimando alla lettura del questionario sottoposto alle unità statistiche.
Essendo ora giunti al termine di questa prima trattazione, il nostro obiettivo è ora crearci un dataset definitivo su cui andare ad eseguire le nostre regressioni multiple.
Ci siamo creati questo dataset alla fine del nostro script, ed esso comprende tutte le variabili che sono a noi utili per l'analisi che andremo a compiere nella seconda parte del progetto.
Rimando, in particolare, alla riga #348 del nostro script, nella quale abbiamo creato il dataset " conviventi_e_sposati". Esso comprende 10 variabili e 7241 valori.

Regressioni lineari multiple

Lo script riguardante le regressioni lineari multiple sarà diviso in due parti: nella prima parte verranno analizzate tre regressioni lineari multiple più generali, mentre nella seconda parte verranno svolte le regressioni lineari multiple dei 20 gruppi più popolosi nel nostro campione,

con il fine di ottenere delle regressioni multiple che riescano a coprire, ponendo che il nostro campione rappresenti fedelmente la popolazione, un'alta percentuale della popolazione (nel nostro caso più del 92% di essa) in maniera molto precisa partendo dall'informazione riguardante l'etnia, la religione, dove vivono e se sono sposati o solo conviventi.
Prima di immergersi nell'analisi vera e propria vorremmo fare un piccolo excursus che vada a definire la caratterizzazione e l'importanza di ogni variabile da noi scelta.
Abbiamo infatti preso in esame nel nuovo dataset "conviventi_e_sposati" 10 variabili diverse, che andando in ordine di posizione nel dataset sono le seguenti:
la prima variabile è la variabile "pid", che classifica ogni unità statistica in base al suo codice personale d'identità composto da 6 cifre. Questo insieme di valori, nonostante possa inizialmente sembrare superfluo ai fini delle nostre analisi quantitative, risulterà di vitale importanza perché ci garantisce l'unicità di ogni unità statistica, in modo tale da evitare eventuali duplicazioni nel modello che andrebbero solamente ad alterare i risultati ottenuti. Abbiamo verificato personalmente l'unicità di ogni singola unità statistica.
La seconda variabile è denominata "popgrp", abbreviazione di population group, che va a suddividere la popolazione del campione in 5 gruppi etnici differenti: 1 = African, 2 = Coloured, 3 = Asian/Indian, 4 = White, 5 = Other.
Successivamente troviamo la variabile dicotomica mar che abbiamo già ampiamente descritto e approfondito nella prima parte del paper, motivo per il quale passiamo immediatamente alla quarta variabile del dataset, ovvero la variabile "rel" abbreviazione di "religious affiliation of respondent". La variabile può assumere 7 diversi stati, che sono: 1= "non religiosi", 2 = "cristiani", 3 = "ebrei", 4 = "musulmani", 5 = "induisti", 6 = "credenze spirituali tradizionali africane", 7 = "altri".
Andando avanti troviamo la variabile "geo2011", che raggruppa la popolazione del campione in 3 differenti nuclei geografici: 1 = coloro che vivono in villaggi, 2 = coloro che abitano in città ed infine 3 = coloro che vivono in campagna. Attraverso il comando: `table(conviventi_e_sposati$geo_2011)` possiamo verificare come la popolazione del nostro campione (7242 unità totali) viva per lo più in contesti urbani (4449 unità) e nei villaggi (2258 unità).
Successivamente è presente la variabile "ppincome", che ci restituisce il reddito mensile familiare diviso il numero delle persone presenti nel nucleo stesso. Alla sua destra, come settima variabile, troviamo "hhsizer" ovvero il numero di componenti del nucleo familiare.
Le ultime tre variabili del dataset ("age", "anni_convivenza", "anni_istruzione") non necessitano di ulteriori spiegazioni (qualora vi fossero dei dubbi, rimando alla prima sezione dello script "regressione logistica e pulizia dei dati" che va ad approfondire riga per riga la corrispondenza degli anni di istruzione con il grado di istruzione specifico).
Come possiamo notare abbiamo volontariamente eliminato la variabile Married che fu di vitale importanza nella prima parte del nostro progetto. Possiamo giustificare la nostra decisione spiegando come, ai fini delle nostre analisi, il secondo esito della variabile dicotomica in questione ("married == 0") risulti inutile, poiché noi siamo interessati a studiare esclusivamente le unità statistiche sposate o conviventi, che vengono classificate attraverso la variabile married nella sua realizzazione 1 (quindi married == 1). Perciò per permettere una più semplice visualizzazione del dataset e non sovraccaricare ulteriormente lo stesso abbiamo scartato questa variabile.
Partendo da queste informazioni possiamo finalmente iniziare il nostro lavoro di analisi statistica andando ad effettuare il primo tentativo di regressione lineare multipla, selezionando come variabile dipendente "anni_convivenza" e come variabili indipendenti

“age”, “hhsizer”, “ppincome”, (la motivazione che ci ha spinto ad inserire solamente tre variabili indipendenti sarà spiegata nella la spiegazione del modello).
Il comando che andremo ad utilizzare da qui in avanti per le regressioni lineari multiple sarà:

Nome_regressione = lm(anni_convivenza ~ “*somma delle variabili indipendenti considerate”****, data=conviventi_e_sposati)***

Dopo aver eseguito la prima regressione sul totale del campione senza distinzione tra le unità conviventi(mar==2) e sposate (mar==1) denominato “regress_multipla_tot” ci possiamo adesso estrapolare le informazioni più importanti attraverso il comando “*summary()”* che andremo, per comodità, ad inserire anche qui sotto nella sua completezza:

```
Residuals:
    Min      1Q  Median      3Q     Max
-48.236  -4.718   0.881   5.756  41.357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.164e+01  3.815e-01 -56.722   <2e-16 ***
age          7.911e-01  6.981e-03 113.320   <2e-16 ***
hhsizer      3.293e-01  3.713e-02   8.869   <2e-16 ***
ppincome     1.202e-05  6.055e-06   1.985   0.0472 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.536 on 7237 degrees of freedom
Multiple R-squared:  0.6457,    Adjusted R-squared:  0.6455
F-statistic:  4396 on 3 and 7237 DF,  p-value: < 2.2e-16
```

Andiamo adesso a commentare i risultati ottenuti.
La prima sezione che ci viene proposta è la sezione "Residuals", che ci restituisce il valore minimo, il massimo, la media, la mediana, primo e terzo quartile.
A seguire troviamo la sezione “Coefficients”, che ci fornisce le stime dei coefficienti per le variabili indipendenti, i loro errori standard, valori t e p-value.
La prima informazione riportata tra parentesi è l’intercetta che da un punto di vista matematico ci suggerisce come la variabile dipendente vada ad assumere il valore -21,6 quando tutte le variabili indipendenti sono nulle. Successivamente vengono quindi riportati nella colonna “Estimate” i coefficienti angolari di ogni variabile indipendente presa in considerazione. Come possiamo notare tutti i coefficienti risultano essere positivi, il che suggerisce una proporzionalità diretta delle variabili indipendenti rispetto alla variabile dipendente “anni_convivenza”.
A seguire troviamo lo “standard error”, che indica la deviazione standard dei residui, ovvero la misura di quanto i dati osservati reali (gli anni reali di matrimonio delle nostre unità statistiche) deviano dai valori predetti dal modello.
L’ultima informazione che si evidenzia nella sezione “Coefficients” è il p-value. Questo valore è di fondamentale importanza nell’analisi di una regressione, poiché definisce quanto la relativa variabile indipendente considerata sia statisticamente significativa (o se non lo sia affatto). Infatti tutte le variabili che assumono un p-value minore di 0.05 risultano essere statisticamente significative per stimare gli anni di convivenza. Sotto questa soglia, minore è il valore e maggiore sarà la significatività. Questo sarà d’ora in avanti la discriminante che useremo nel decidere quali variabili considerare significative. Prima di ogni regressione

infatti andremo a verificare il valore del p-value per ogni variabile indipendente, e qualora esso sia maggiore di 0.05 andremo automaticamente a scartare la relativa variabile.
A seguire notiamo R-squared e adjusted R-squared, due metriche utilizzate per valutare la qualità di un modello di regressione lineare. In questo caso il nostro modello e le sue variabili indipendenti sono in grado di spiegare al 64% la variazione della variabile dipendente. La differenza sostanziale tra R-squared ed adjusted R-squared è che il primo tende ad aumentare quando introduciamo ulteriori variabili indipendenti al modello, anche se esse non sono significative. Ciò può portare ad una sovrastima della bontà del modello quando si includono molte variabili, mentre adjusted R-squared è un indicatore più affidabile. Tuttavia nel nostro caso la differenza tra i due parametri è quasi inesistente, e questo è dovuto al fatto che nel nostro modello stesso abbiamo già eliminato le variabili non significative.
Conclusa questa introduzione generale sui modelli di regressione lineare multipla, passiamo ora all'effettivo confronto delle varie classi della popolazione, in modo tale da ricavarne i giusti spunti di riflessione.
Dopo aver approfondito la regressione multipla comprendente tutta la popolazione del campione, ci siamo dunque interrogati su quali potessero essere le differenze sostanziali tra i due macro gruppi principali (sposati e conviventi). Dopo aver generato i due modelli abbiamo nuovamente estrapolato le summary che riporteremo qui sotto per un veloce confronto:

regress_multipla_sposati

```
Residuals:
    Min      1Q  Median      3Q     Max
-49.973  -4.799   1.115   6.167  41.708

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.737962   0.506525  -46.86  < 2e-16 ***
age           0.843415   0.008867   95.12  < 2e-16 ***
hhsizer       0.274586   0.045459    6.04 1.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.823 on 5167 degrees of freedom
Multiple R-squared:  0.641,     Adjusted R-squared:  0.6408
F-statistic:  4613 on 2 and 5167 DF,  p-value: < 2.2e-16
```

regress_multipla_conviventi

```
Residuals:
     Min       1Q   Median       3Q      Max
-29.6299  -4.1585  -0.1899   3.6064  29.6803

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.11622    0.89393 -10.198  < 2e-16 ***
age              0.53809    0.01481  36.345  < 2e-16 ***
anni_istruzione -0.19149    0.05018  -3.816  0.00014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.196 on 2068 degrees of freedom
Multiple R-squared:  0.4713,    Adjusted R-squared:  0.4708
F-statistic: 921.6 on 2 and 2068 DF,  p-value: < 2.2e-16
```

La prima cosa che possiamo notare è che l'unica variabile indipendente statisticamente rilevante per entrambi i modelli è la variabile "age", che oltre a differire di qualche punto decimale, risulta essere positivamente correlata con gli anni di convivenza in entrambi i casi. Questo è naturalmente dovuto al fatto che, più è anziana l'unità statistica, e più è probabile che la durata della convivenza o del matrimonio aumenti.
Per gli individui sposati, la seconda variabile statisticamente rilevante è "hhsizer", che risulta avere anch'essa un coefficiente positivo. Da ciò si deduce che gli anni di convivenza con il proprio partner, se si è formalmente sposati, sono positivamente proporzionali alla dimensione del nucleo familiare, e quindi si tenderà, in una famiglia, ad essere più uniti qualora si abbiano figli o genitori a carico.
La seconda variabile statisticamente significativa per gli individui conviventi è invece "anni_istruzione", Possiamo ora notare che il coefficiente, a differenza delle variabili appena descritte, risulta qui essere negativo, evidenziando come gli anni di convivenza siano inversamente proporzionali agli anni di istruzione dell'unità statistica. Il coefficiente negativo qui evidenziato ci ha portato, dopo aver notato come esso sia ricorrente quando legato alla variabile indipendente anni_istruzione a riflettere su questo fenomeno. Ciò che ne è venuto fuori verrà esposto più nel dettaglio più avanti.

Potremmo ora studiare come cambia la regressione multipla qualora si impongano alla stessa dei vincoli diversi da quelli imposti nel caso delle due regressioni precedenti (l'essere sposati o l'essere solo conviventi), quali ad esempio la religione, l'etnia e il luogo di residenza.
Nonostante il sicuro interesse che tali regressioni avrebbero potuto suscitare nel lettore, abbiamo deciso qui di dividere la nostra popolazione in diversi gruppi, combinando tra di esse le variabili rel, mar, popgrp e geo2011, ed in base ad esse ci siamo poi andati ad estrapolare i 20 gruppi più numerosi presenti nella popolazione. Ad esempio nel primo gruppo rientrano tutti coloro che sono sposati, neri, cristiani vivono in città. Nel nostro campione essi arrivano ad essere un totale di 1385 unità statistiche, pari al 19.127% della popolazione totale. Il ventunesimo gruppo, invece, contiene tutte quelle unità statistiche che non possono essere inserite in nessuno dei precedenti gruppi, nonostante rientrino in conviventi_e_sposati. Queste unità risultano essere 545, ossia il 7.527% delle unità totali.
Per ulteriori informazioni riguardanti i singoli gruppi rimandiamo alla lettura dello script. Affinché però la lettura dello script stesso risulti semplice e agevole, sottolineo i seguenti passaggi.
All'inizio della disamina di ogni gruppo abbiamo inserito il numero di unità che fanno parte del gruppo in questione, e grazie ad esso abbiamo poi calcolato la percentuale di unità che ricadevano in esso.
Abbiamo poi effettuato la regressione lineare multipla, andando dapprima ad includere tutte le quattro variabili indipendenti da noi prese in esame e poi andando ad eliminare quelle non statisticamente rilevanti. Ci siamo poi andati a creare una variabile r quadro, che utilizzeremo alla fine dello script.
Abbiamo poi creato il "Potential Group", ossia un gruppo contenente tutte le potenziali unità statistiche del dataset_pulito a cui il singolo modello di regressione è applicabile. In altre parole, come già accennato in precedenza, qualora vi fossero delle variabili indipendenti statisticamente non significative, nel potential group andremmo ad inserire anche quelle unità che, in presenza della variabile non significativa, presentano un valore na. Questo è possibile farlo perché la variabile non è statisticamente significativa.
Alla fine dello script abbiamo anche inserito una somma delle unità totali analizzabili come somma dei potential group, che risulta essere 7378.
Inoltre abbiamo applicato in ogni gruppo la funzione *predict*, che ci ha permesso di stimare gli anni di convivenza per tutte quelle unità rientranti nel potential group. Abbiamo successivamente aggiunto questi valori nella colonna "anni_mar_stimato", che è stata poi utilizzata alla fine dello script, alla riga #979, per integrare il dato riguardante gli anni di convivenza o di matrimonio qualora alcune unità non lo avessero disponibile (per poter poi eseguire in maniera più precisa il calcolo dell'intervallo di confidenza).
Queste unità aggiuntive ci serviranno più avanti, quando andremo ad utilizzare direttamente il dataset pulito per effettuare gli intervalli di confidenza.
Per una più semplice interpretazione dei gruppi, da parte del lettore, abbiamo inoltre qui sotto specificato come questi 21 gruppi sono composti.
Abbiamo inoltre assunto che il nostro campione sia un campione che rispecchia perfettamente l'intera popolazione, e andremo quindi ad esplicitare anche la percentuale di popolazione che ricade in quel gruppo.

Gruppo 1: cristiani neri sposati che abitano in città. 19.127% della popolazione.
Gruppo 2: cristiani neri sposati che vivono in villaggi. 17.746% della popolazione.
Gruppo 3: cristiani mulatti sposati che vivono in città. 10.33% della popolazione.

Gruppo 4: cristiani neri conviventi che vivono in città. 9.253% della popolazione.
Gruppo 5: cristiani bianchi sposati che vivono in città. 8.162% della popolazione.
Gruppo 6: cristiani neri conviventi che vivono in villaggi. 6.049% della popolazione.
Gruppo 7: cristiani mulatti conviventi che vivono in città. 4.267% della popolazione.
Gruppo 8: coloro che hanno credenze spirituali tradizionali africane neri sposati che vivono in villaggi. 2.859% della popolazione.
Gruppo 9: non religiosi neri sposati che vivono in villaggi. 1.809% della popolazione.
Gruppo 10: cristiani neri sposati che vivono in campagna. 1.547% della popolazione.
Gruppo 11: cristiani mulatti sposati che vivono in campagna. 1.409% della popolazione.
Gruppo 12: coloro che hanno credenze spirituali tradizionali africane neri sposati che vivono in città. 1.395% della popolazione.
Gruppo 13: cristiani neri conviventi che vivono in campagna. 1.395% della popolazione.
Gruppo 14: non religiosi neri conviventi che vivono in villaggi. 1.201% della popolazione.
Gruppo 15: non religiosi neri conviventi che vivono in città. 1.146% della popolazione.
Gruppo 16: non religiosi neri sposati che vivono in città. 1.063% della popolazione.
Gruppo 17: coloro che hanno credenze spirituali tradizionali africane neri conviventi che vivono in villaggi. 1.049% della popolazione.
Gruppo 18:coloro che hanno credenze spirituali tradizionali africane neri conviventi che vivono in città. 0.967% della popolazione.
Gruppo 19: cristiani mulatti conviventi che vivono in campagna. 0.911% della popolazione.
Gruppo 20: induisti asiatici/indiani sposati che vivono in città. 0.787% della popolazione.
Gruppo 21: resto della popolazione. 7.527% della popolazione.

Per facilitare al meglio la comprensione della distribuzione della popolazione abbiamo deciso di inserire inoltre questo grafico a torta, per avere graficamente un'idea di come la popolazione sia distribuita.
Andiamo ora infine a commentare i risultati ottenuti dai 21 modelli per trarre le dovute conclusioni. Ovviamente non analizzeremo ogni singolo output dei 21 modelli (confidando anche nella capacità interpretativa del lettore), ma proporremo delle riflessioni generali che abbiamo tratto da essi.
Prima di partire è inoltre utile menzionare che abbiamo calcolato la media degli r-quadro, in modo tale da avere bene in mente quanto fedelmente questi modelli descrivono la realtà.
R quadro medio risulta essere 0.53, che sebbene non sia un valore altissimo è comunque un valore che, ai fini della nostra analisi, è più che accettabile.
In particolare abbiamo che, qualora l'età di convivenza o matrimonio aumenti di un anno, questo sarà dovuto mediamente per il 53% dalle nostre quattro variabili numeriche prese in esame nella creazione delle regressioni multiple di cui sopra (income, dimensione del nucleo familiare, età e anni istruzione) e per un 47% da altri fattori da noi non analizzati. Per effettuare questa analisi abbiamo ricorso ad una media ponderata, e per ulteriori spiegazioni rimando alla riga #938 dello script.
Passando alle riflessioni vere e proprie, la prima cosa che possiamo estrapolare dalle regressioni da noi create è che, ogni qual volta la variabile anni_istruzione risulti significativa, essa è legata ad anni_convivenza da un coefficiente negativo.
Abbiamo cercato delle spiegazioni a questo fenomeno, e le più importanti che abbiamo individuato sono le seguenti.
La prima spiegazione è che probabilmente le relazioni più durature si vengono a creare negli anni di gioventù, che sono i più indicati per questa attività (l'andarsi a creare relazioni

durature). Tuttavia, se in questo periodo l'individuo tende a concentrarsi maggiormente nella propria formazione, avrà meno tempo e chance per costruire una solida relazione con il proprio partner.
Una seconda spiegazione plausibile è che, se in gioventù l'individuo ha deciso di studiare, è più probabile che si trovi un partner dopo aver concluso gli studi, e quindi avrà meno tempo a disposizione per far durare un eventuale convivenza oppure un eventuale matrimonio (perchè appunto ha trovato il partner ad un'età più avanzata rispetto a coloro che non studiano). Viceversa, un individuo che ha scelto di non studiare si ritroverà probabilmente ad intraprendere una relazione duratura con il proprio partner già nel periodo che avrebbe dovuto/potuto dedicare allo studio, sposandosi probabilmente prima e quindi facendo durare il matrimonio più a lungo.
Inoltre possiamo notare come l'effetto maggiore si abbia sulle convivenze e non sui matrimoni, questo perché è più probabile che nel periodo di gioventù, nel quale si tende a studiare, la coppia conviva piuttosto che sia sposata. Avendo ora, gli individui conviventi, un legame più debole con il proprio partner, la coppia farà più difficoltà a "superare" il periodo in cui uno dei due individui ha maggiori impegni esterni alla relazione (come è, ad esempio, lo studio), ed è più probabile quindi che la convivenza si interrompa. Per gli individui sposati, dall'altro lato, avendo un legame più forte, sarà più semplice superare le difficoltà, e quindi un eventuale periodo di studio di uno dei due individui della coppia non andrà ad influire così tanto come avrebbe influito nel caso dei conviventi.
Ultima osservazione, piuttosto banale e dovuta al modo con cui sono stati raccolti i dati di cui disponiamo, è che qualora si avesse abitato al di fuori del nucleo familiare nel quale risiede il partner (pensiamo agli studenti fuori sede oppure agli studenti che abbiano svolto un periodo di studio all'estero) la convivenza risulta essere interrotta, e si inizia a conteggiare il periodo di convivenza partendo nuovamente da zero una volta che l'individuo in questione è tornato a convivere. Nel caso invece l'individuo sia sposato e non abitasse nel nucleo familiare dove risiede sua moglie per un certo periodo di tempo, quest'ultimo periodo verrebbe comunque conteggiato nella durata del matrimonio.
Ultimo punto su cui vorremmo porre l'attenzione riguarda la regressione multipla del gruppo 18. Qualora infatti la variabile anni di istruzione risulti statisticamente rilevante nei nostri modelli, questa variabile presenta un coefficiente negativo. Unica eccezione è la regressione del gruppo 18, in cui "anni_istruzione" risulta avere un coefficiente positivo di 0.44. Andando però ad analizzare meglio la regressione di questo gruppo notiamo che, nonostante il p-value sia sotto la soglia dello 0.05%, facendo quindi risultare la variabile in questione come statisticamente rilevante, esso sia in realtà pari a 0.0496, valore al limite della rilevanza statistica. Per questo motivo, ai fini della nostra analisi interpretativa, abbiamo deciso di non prendere in considerazione il diciottesimo gruppo nello sviluppo delle nostre riflessioni riguardo la variabile anni_istruzione.
Nel terzo ed ultimo script ci siamo infine andati a cercare l'intervallo di confidenza, per inquadrare la media degli anni di convivenza della popolazione partendo dalla media del nostro campione. Per questa operazione possiamo utilizzare il nostro "dataset_pulito", che abbiamo avuto modo di modificare alla fine del secondo script integrando, ove possibile, i dati mancanti della colonna anni_conviventi creando la colonna anni_conv_matrim. Abbiamo calcolato gli intervalli di confidenza sia per gli individui sposati che per gli individui conviventi, ed abbiamo imposto due differenti livelli di significatività. Nel calcolare il primo intervallo, abbiamo imposto l'alfa pari a 0.05 e ci siamo ricavati il limite inferiore e superiore del nostro intervallo di confidenza, prima degli sposati, con una media compresa tra i 20 anni e 6 mesi e 21 anni e 3 mesi, poi dei conviventi, con una media della popolazione compresa tra i 10

anni e 4 mesi e gli 11 anni (è quindi probabile al 99.5% che questi due intervalli comprendano la durata del matrimonio e della convivenza dell'intera popolazione). Il secondo passo è stato quello di imporre un livello di significatività molto più alto (creando quindi un intervallo più preciso) ponendo alfa pari a 0.00001.
Nel caso degli sposati, l'intervallo è risultato essere quello compreso tra 20 anni e 2 mesi e 21 anni e 9 mesi, mentre nel caso dei conviventi l'intervallo diviene quello compreso tra 9 anni ed 8 mesi e 11 anni e 7 mesi. I risultati ottenuti sono stati anche riportati direttamente sul nostro script.

RINGRAZIAMENTI

Concludendo questo nostro elaborato, cogliamo l'occasione per porgere i seguenti ringraziamenti.
Per prima cosa, vorremmo ringraziare la professoressa Paola Vicard, che non solo ci ha formato l'anno passato, con il corso di statistica base, ma ha fornito a noi studenti anche delle solide basi per affrontare quest'anno il corso di statistica per big data. Ne ricordiamo in particolare la passione per la materia che ha animato e contraddistinto ogni sua lezione.
Vorremmo poi ringraziare il tutor di quest'anno per quel che riguarda il corso di statistica per big data, Flavio Mangione, che ha sempre fatto il massimo per venire incontro alle esigenze di noi studenti.
Infine, vorremmo porgere i nostri ringraziamenti più sentiti al Chiarissimo Professor Francesco Dotto, il quale ha tenuto il corso di statistica per big data e che si è sempre saputo distinguere per l'intrinseca passione per la statistica, che ha sempre cercato di trasmettere, nel caso di noi tre con successo, ai suoi studenti. Il Professor Dotto è stato inoltre il primo che, nel nostro percorso universitario, ci ha fatto lavorare autonomamente sui dati, dando un assetto anche di tipo pratico alle sue lezioni (ricordo in particolare le sue lezioni su r il venerdì) e al suo corso. Grazie a questo approccio di tipo anche pratico, le sue lezioni si sono sempre rivelate interessanti e di facile comprensione. Lo ricordiamo e lo ringraziamo inoltre per la disponibilità che, per tutta la durata del corso, ha dimostrato di avere nei confronti di noi studenti, essendo sempre pronto a chiarire, sia a lezione sia dopo le lezioni o ai ricevimenti, ogni nostro eventuale dubbio.