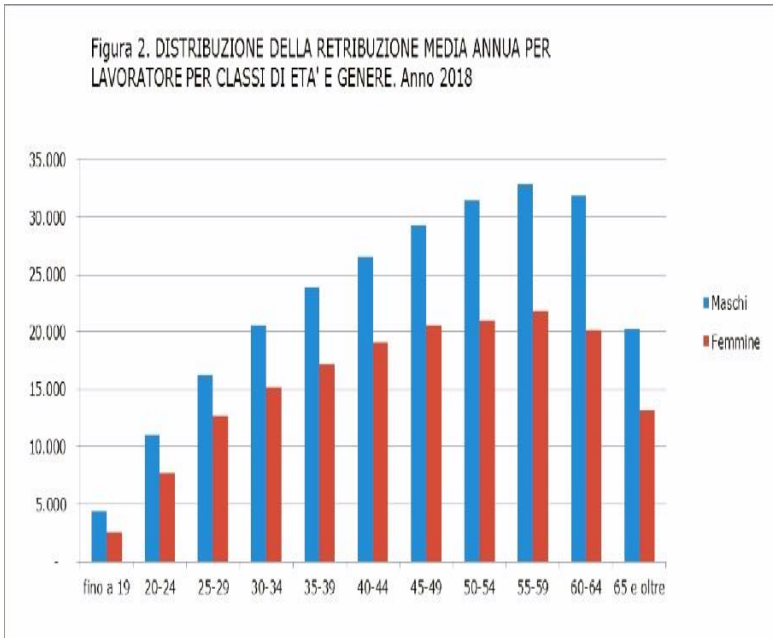
Tesina sulla Retribuzione Media Annua dei Lavoratori Dipendenti

Indice

1. Introduzione

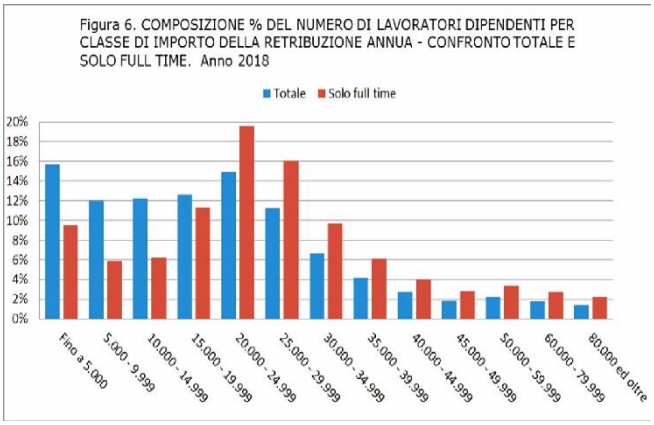
La retribuzione media annua dei lavoratori dipendenti è il rapporto tra la retribuzione totale annua (al lordo Irpef) dei lavoratori dipendenti del settore privato non agricolo assicurati presso l'Inps e il numero dei lavoratori dipendenti, la quale viene calcolata in euro. È una variabile osservabile negli indicatori BES (Benessere Equo e Sostenibile), un insieme di indicatori cui compito è di valutare il progresso della società sia dal punto di vista economico che sotto l'aspetto sociale e ambientale.

La nostra analisi si soffermerà sullo stabilire quali sono gli aspetti di maggiore importanza nella determinazione della retribuzione media annua e come essi impattano sullo spettro dell'area geografica nazionale.



Come possiamo affermare dai dati forniti dall'Osservatorio sul lavoro dipendente privato (Coordinamento statistico-attuariale dell'INPS) del 2018, la retribuzione media è crescente all'aumentare dell'età del lavoratore, fino ad una soglia di 60 anni. Ciò accade in maniera eguale se guardiamo alla differenza di retribuzione tra uomo e donna.

Un altro dato interessante è dato dalla percentuale di lavoratori suddivisa per classi di importo. Più le cifre si alzano, più si scopre che ad ottenerle sono i lavoratori full-time, mentre mediamente un lavoratore part-time riscuote importi minori, che potrebbero derivare da un tempo lavorativo limitato all'interno dell'arco annuale considerato.



2. Obiettivi

Il nostro compito partirà dall'analisi di un data-set che fa riferimento all'appendice statistica del rapporto BES dell'ISTAT, che offre un quadro integrato dei principali fenomeni economici, ambientali e sociali riguardanti il nostro Paese, attraverso l'analisi di un set di indicatori. Esso si compone di 87 osservazioni e di 53 variabili:

Le prime tre sono di natura qualitativa:

- Regioni
- Province
- Ripartizione geografica

Le altre 50 sono suddivise in 11 macrocategorie:

1) Ambiente
2) Benessere economico
3) Innovazione-ricerca e creatività
4) Istruzione e formazione
5) Lavoro e conciliazione dei tempi di vita
6) Paesaggio e patrimonio culturale
7) Politica e istituzioni
8) Qualità dei servizi
9) Relazioni sociali
10) Salute
11) Sicurezza

Il nostro primo passo della suddetta analisi è stato effettuare una selezione dei dati, concentrandoci sulle caratteristiche di base delle nostre variabili, ed in particolare sulla loro correlazione con la variabile dipendente da noi scelta. Dopo questo passaggio, il nostro secondo punto è stato quello di analizzare in maniera descrittiva le nostre variabili scelte tramite l'utilizzo di grafici, in modo da indagare sulla natura di ogni fenomeno ed osservare se esso sia uniforme o meno in tutto il nostro paese. La terza fase consiste nella stima di un modello di regressione lineare multiplo che metta in evidenza le variabili che più concorrono alla formazione della retribuzione media annua, sia in modo positivo, che in quello negativo, delineando una puntualizzazione di base del modello con la stima dei parametri e i due indici R-quadro riguardanti la bontà di adattamento del modello ai dati. Successivamente vengono individuati gli intervalli di confidenza dei regressori e verificandone le loro ipotesi di significatività, verranno selezionati con delle procedure di stepwise automatiche che si basano sull'indice AIC (Akaike Information

Criterion), per poi arrivare ad una specificazione finale del modello, analizzandone i segni dei coefficienti di regressione, la loro ammissibilità e la verifica delle ipotesi alla base del modello.
Infine, si effettua una regressione multipla di tipo logit, questo tipo di regressione può essere effettuata avendo la nostra variabile dipendente di tipo binario; fatto ciò, vengono utilizzati una serie di regressori selezionati tramite la stepwise basata sull' AIC, per poi riportare la curva ROC (Receiver operating characteristic) e l'indice AUC (Area Under the Curve), i quali verificano rispettivamente la bontà e la correttezza del modello stimato.

3. Presentazione del dataset e selezione dei dati

La prima fase consiste nell'analizzare le variabili di riferimento, evidenziando per ognuna di esse le proprie statistiche basilari, per capire se esse potranno esserci utili per la nostra analisi o meno, ed in quest'ultimo caso ci occuperemo di ometterle.
I dati che utilizziamo per tale operazione sono quelli che R ci offre in output con il comando summary(), quali: valore minimo, primo quartile, mediana, media, terzo quartile e valore massimo. Inoltre, abbiamo deciso di omettere le righe contenenti dei dati NA (Not Available), per evitare complicazioni.

Ambiente

**(Alessandria, Avellino, Bergamo, Brescia; Como, Lodi, Messina, Monza-Brianza, Padova, Pavia; Rovigo, Torino, Treviso, Verona, Vicenza)

Ambiente							
	Conferimento dei rifiuti urbani in discarica	Disponibilità di verde urbano	Energia elettrica da fonti rinnovabili	Impermeabilizzazione del suolo da copertura artificiale	Qualità dell'aria urbana - Biossido di azoto	Qualità dell'aria urbana - PM10	Raccolta differenziata dei rifiuti urbani
Minimo	0.00	3.60 (Crotone)	3.70 (Firenze)	2.70 (Matera)	0.000	0.00	19.90 (Palermo)
1° Quartile	0.00	15.65	13.80	6.00	0.000	0.00	19.90
Mediana	9.40	23.30	26.10	8.10	0.000	0.00	48.10
Media	28.04	54.08	54.89	9.35	9.068	23.44	62.30
3° Quartile	37.15	40.65	54.60	11.40	0.000	50.00	59.43
Massimo	538.90 (Crotone)	997.20 (Matera)	497.80 (Sondrio)	34.00 (Monza-Br.	100.000 Como; MI	100.00 **	87.30 (Treviso)

Riguardo alla macro-area ambiente, abbiamo deciso di tralasciare tutte le possibili variabili, in quanto le abbiamo ritenute non pertinenti rispetto alla nostra variabile dipendente.

Benessere economico

Benessere economico						
	Importo medio annuo pro-capite dei redditi pensionistici	Patrimonio pro capite	Pensionati con pensione di basso importo	Reddito medio disponibile pro capite	Retribuzione media annua dei lavoratori dipendenti	Tasso di ingresso in sofferenza dei prestiti bancari alle famiglie
Minimo	14127 (Crotone)	76489 (V.Valentia)	6.100 (Biella)	10881 (Crotone)	13376 (V.Valentia)	0.600 (Bolzano)
1° Quartile	17056	116521	7.400	14836	16892	0.800
Mediana	18425	164767	8.300	19247	19761	1.100
Media	18278	157031	9.866	18229	19778	1.074
3° Quartile	19383	187747	11.950	20588	22749	1.300
Massimo	22573 (Milano)	295154 (Milano)	18.800 (Crotone)	27301 (Milano)	30092 (Milano)	1.800 (Catania)

Riguardo al benessere economico:
Qui abbiamo avuto le prime correlazioni con la nostra variabile dipendente: le variabili “Importo medio annuo pro capite dei redditi pensionistici” e “Pensionati con pensione di basso reddito”, hanno registrato un’alta correlazione, ma per evitare multicollinearità, abbiamo deciso di prendere in considerazione solo la prima, in quanto è risultata con l’indice di correlazione maggiore (0.84).
Decidiamo, successivamente, di considerare nella nostra analisi le variabili “Patrimonio pro capite” e “Reddito medio disponibile pro capite” in quanto ritenute pertinenti con la nostra analisi e con una correlazione alta, rispettivamente pari a 0.84 e 0.92.
La variabile “Tasso di ingresso in sofferenza dei prestiti bancari alle famiglie” non è stata presa in considerazione.

Innovazione, ricerca e creatività

Innovazione, ricerca e Creatività		
	Addetti nelle imprese culturali	Mobilità dei laureati italiani (25-39 anni
Minimo	0.60 (Taranto)	-58.500 (Crotone)
1° Quartile	0.900	-15.500
Mediana	1.100	-7.100
Media	1.133	-8.839
3° Quartile	1.200	.-0.15
Massimo	3.000 (Milano)	38.400 (Bologna)

Riguardo all'innovazione, ricerca e creatività, abbiamo deciso di considerare esclusivamente la variabile "Mobilità dei laureati italiani (25-39 anni)", che ha una correlazione di 0.76, in quanto la mobilità è da noi considerata come un fattore importante per la determinazione della nostra variabile dipendente, visto anche gli squilibri economici all'interno del nostro paese pendono tendenzialmente a favore del Nord rispetto alle regioni meridionali.

Istruzione e formazione

Istruzione e formazione							
	Competenza alfabetica non adeguata	Competenza numerica non adeguata	Giovani che non lavorano e non studiano (NEET)	Laureati e altri titoli terziari (25-39 anni)	Partecipazione alla formazione continua	Passaggio all'università	Persone con almeno il diploma (25-64 anni)
Minimo	19.30 (Trento)	18.10 (Trento)	9.70 (Pordenone)	12.00 (Crotone)	3.500 (Foggia)	37.10 (Sondrio)	43.00 (Bar-And-Tr)
1° Quartile	26.95	31.40	15.05	23.05	6.100	47.85	54.75
Mediana	32.20	37.30	18.40	26.00	7.400	51.50	62.30
Media	33.67	39.94	21.08	26.42	7.841	51.03	60.53
3° Quartile	39.35	47.35	25.30	30.65	9.300	54.42	66.30
Massimo	59.20 (Enna)	69.60 (Crotone)	48.20 (Caltaniss.)	43.80 (Bologna)	15.900 (Bologna)	62.60 (Isernia)	74.90 (Bologna)

Riguardo a Istruzione e Formazione:
Abbiamo deciso di considerare la variabile “Competenza numerica non adeguata” e di escludere la variabile “Competenza alfabetica non adeguata”, in quanto la prima ha un coefficiente di correlazione negativo maggiore rispetto alla seconda, pari a -0.78.
Inoltre, abbiamo ritenuto necessario considerare anche le variabili “Giovani che non lavorano e non studiano” e “Persone con almeno il diploma”, con correlazioni rispettivamente pari a -0.75 e 0.64.

Lavoro e conciliazione dei tempi di vita

Lavoro e conciliazione dei tempi di vita						
	Giornate retribuite nell'anno (lavoratori dipendenti)	Tasso di infortuni mortali e inabilità permanente	Tasso di mancata partecipazione al lavoro	Tasso di mancata partecipazione al lavoro giovanile (15-29 anni)	Tasso di occupazione (20-64 anni)	Tasso di occupazione giovanile (15-29 anni)
Minimo	64.60 (Salerno)	4.80 (Biella)	6.70 (Belluno)	12.90 (Belluno)	42.00 (Trapani)	14.90 (Benevento)
1° Quartile	72.25	10.00	10.90	21.70	59.10	28.15
Mediana	77.70	12.20	13.20	27.40	69.20	35.20
Media	76.50	13.03	18.03	33.36	64.72	33.08
3° Quartile	81.55	15.05	22.90	44.35	72.30	39.30
Massimo	85.50 (Lecco)	29.90 (Arezzo)	44.60 (Trapani)	70.10 (Trapani)	77.00 (Salerno)	44.90 (Brescia)

Riguardo al lavoro e conciliazione dei tempi di vita:
Abbiamo ritenuto fondamentale includere nell’analisi la variabile “Giornate retribuite nell’anno (lavoratori dipendenti)”, poiché la riteniamo come uno dei punti cardine del nostro lavoro, visto che è strettamente proporzionale alla nostra variabile dipendente con un’alta correlazione, pari a 0.92.
Il “tasso di infortuni mortali e inabilità permanente” è stato ritenuto ininfluente sulla retribuzione media annua.
Il “Tasso di mancata partecipazione al lavoro” è stato ritenuto congruente con la nostra variabile dipendente e, per evitare multicollinearità, abbiamo deciso di escludere il “Tasso di mancata partecipazione al lavoro giovanile (15-29 anni)”, in quanto la nostra analisi non contempla un range così ristretto di età da analizzare.
Lo stesso discorso è valso con il “Tasso di occupazione (20-64)” ed il “Tasso di

occupazione giovanile (15-29 anni)", includendo il primo ed escludendo il secondo, in quanto meno ampio.

Paesaggio e patrimonio culturale

	Paesaggio e patrimonio culturale		
	Densità di verde storico	Densità e rilevanza del patrimonio museale	Diffusione delle aziende agrituristiche
Minimo	0.000 (Biella-Oristano-Kr)	0.100 (Lodi)	1.200 (Reggio Calabria)
1° Quartile	0.2000	0.250	4.050
Mediana	0.7000	0.700	6.600
Media	1.587	2.239	8.056
3° Quartile	1.750	1.200	10.050
Massimo	19.500 (Pordenone)	40.500 (Napoli)	30.000 (Siena)

Riguardo al paesaggio e patrimonio culturale:
Non abbiamo ritenuto che alcuna di queste variabili potesse essere rilevante con la nostra variabile dipendente

Politica e Istituzioni

	Politica e Istituzioni					
	Affollamento degli istituti di pena	Amministratori comunali con meno di 40 anni	Amministratori comunali donne	Amministrazioni provinciali: capacità di riscossione	Comuni: capacità di riscossione	Partecipazione elettorale
Minimo	0.0 (Sv.-Mc.	13.40 (Trieste)	22.50 (Catanzaro)	29.00 (Trieste)	66.80 (Trento)	34.70 (S.Sardegna)
1° Quartile	107.2	25.55	29.10	65.60	76.60	52.50
Mediana	124.8	28.50	32.50	78.70	78.50	61.30
Media	123.7	28.32	32.17	74.37	78.13	58.75
3° Quartile	144.1	31.40	35.00	86.30	80.40	65.90
Massimo	196.7 (Taranto)	43.60 (Crotone)	40.20 (Bologna)	97.50 (Sondrio)	84.30 (Potenza)	70.20 (Firenze)

Riguardo a Politica e Istituzioni:
non abbiamo ritenuto alcuna variabile utile per la nostra analisi.

Qualità dei servizi

Qualità dei servizi				
	Bambini che hanno usufruito dei servizi comunali per l'infanzia	Emigrazione ospedaliera in altra regione	Irregolarità del servizio elettrico	Posti-km offerti dal Tpl
Minimo	1.00 (Crotone)	1.400 (Lecco)	0.500 (Trieste)	299.3 (Ragusa)
1° Quartile	7.70	4.100	1.400	1283.4
Mediana	12.70	7.000	1.700	1976.6
Media	14.04	8.117	2.074	2568.6
3° Quartile	18.55	10.800	2.450	3312.7
Massimo	32.20 (Bologna)	24.900 (Campob.)	5.500 (Trapani)	15272.0 (Milano)

Riguardo alla qualità dei servizi:
non abbiamo ritenuto alcuna variabile utile per la nostra analisi.

Relazioni sociali

Relazioni sociali		
	Ogranizzazioni non profit	Scuole accessibili
Minimo	31.90 (Napoli)	18.90 (Belluno)
1° Quartile	54.90	30.10
Mediana	65.50	36.20
Media	64.43	36.03
3° Quartile	72.30	40.65
Massimo	116.00 (Agrigento)	66.80 (Aosta)

Riguardo alle relazioni sociali:
non abbiamo ritenuto alcuna variabile utile per la nostra analisi.

Salute

Salute					
	Mortalità infantile	Mortalità per demenze e malattie del sistema nervoso (65 anni e più)	Mortalità per incidenti stradali (15-34 anni)	Mortalità per tumore (20-64 anni)	Speranza di vita alla nascita
Minimo	0.700 (Im.-La Sp.)	22.50 (Matera)	0.0000 (Isernia)	6.60 (Grosseto)	81.10 (Napoli)
1° Quartile	1.700	29.15	0.4500	7.80	82.55
Mediana	2.200	33.40	0.6000	8.30	83.00
Media	2.492	33.83	0.7379	8.39	80.03
3° Quartile	3.050	37.50	0.9500	8.90	83.50
Massimo	5.800 (Trapani)	60.50 (Aosta)	27.000 (Sondrio)	11.30 (Sondrio)	84.40 (Firenze)

Riguardo alla Salute:
non abbiamo ritenuto alcuna variabile utile per la nostra analisi.

Sicurezza

Sicurezza				
	Altri delitti violenti denunciati	Delitti diffusi denunciati	Mortalità stradale in ambito extraurbano	Omicidi
Minimo	8.40 (Udine)	57.8 (Potenza)	1.00 (Rimini)	0.0000 (Aosta)
1° Quartile	12.25	117.2	3.10	0.1500
Mediana	14.10	144.7	4.30	0.4000
Media	15.00	162.0	4.70	0.5057
3° Quartile	17.10	194.1	5.25	0.7500
Massimo	30.10 (Napoli)	411.8 (Milano)	16.10 (Vercelli)	2.300 (Crotone)

Riguardo alla Sicurezza: non abbiamo ritenuto alcuna variabile utile per la nostra analisi.

3.1 Risultati della selezione

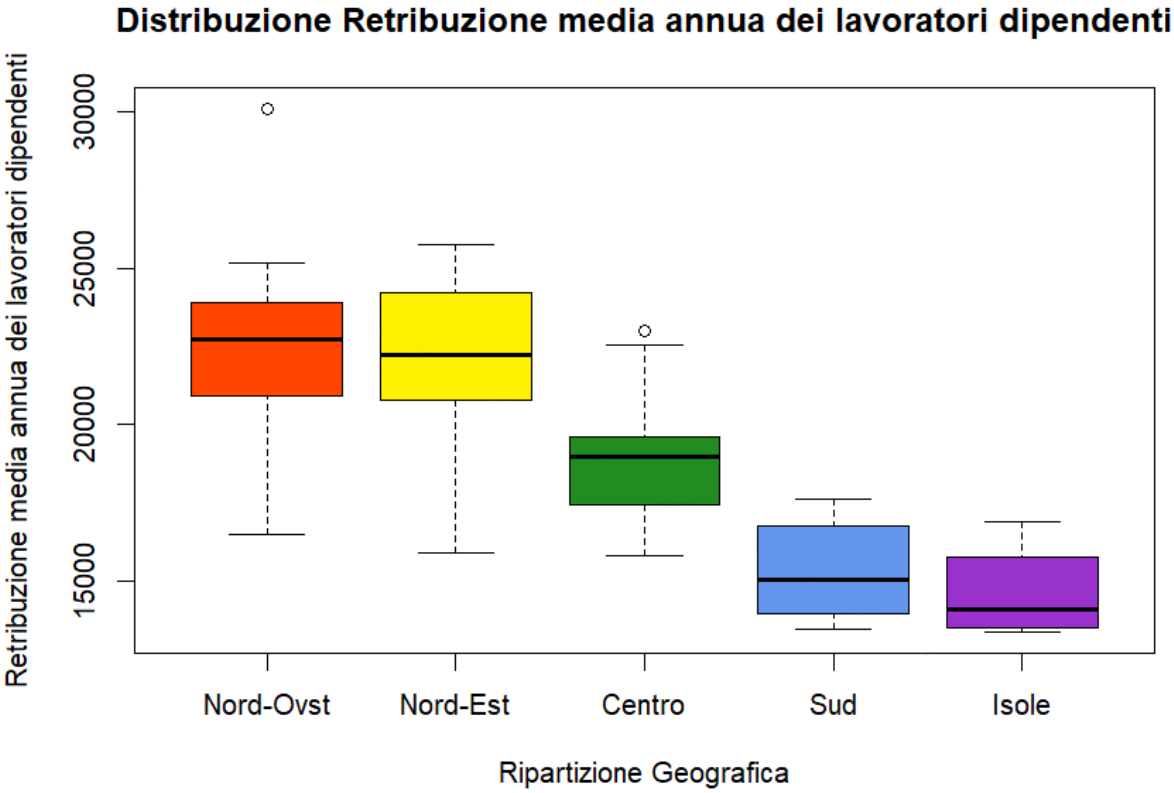
Le variabili indipendenti scelte per sviluppare la nostra regressione multipla sono:

- Importo medio annuo pro capite dei redditi pensionistici
- Patrimonio pro capite
- Reddito medio disponibile pro capite
- Mobilità dei laureati italiani (25-39 anni)
- Competenza numerica non adeguata
- Giovani che non lavorano e non studiano (NEET)
- Persone con almeno il diploma
- Giornate retributive nell'anno (lavoratori dipendenti)
- Tasso di mancata partecipazione al lavoro
- Tasso di occupazione (20-64 anni)

4. Analisi descrittiva dei dati

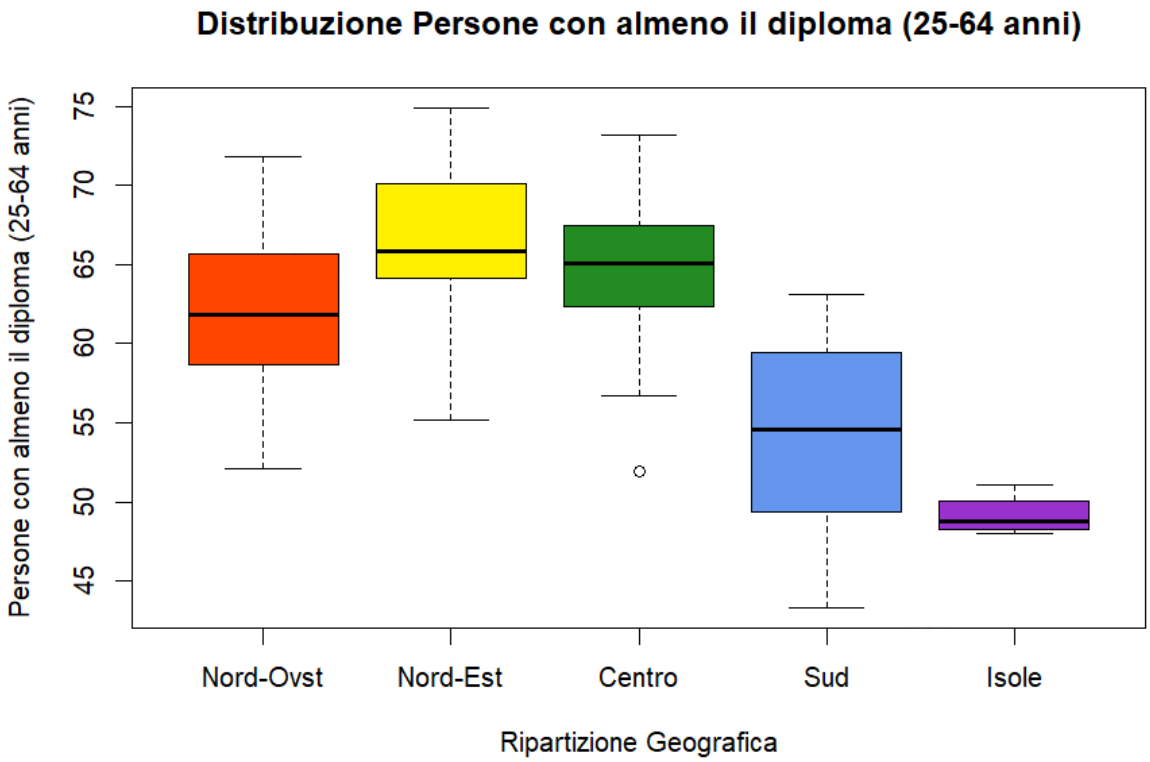
Prima di analizzare i dati delle variabili rimaste dopo la selezione dei dati, abbiamo suddiviso per ripartizione geografica le regioni, per osservare dove i nostri fenomeni sono più o meno osservabili.

Retribuzione media annua dei lavoratori dipendenti



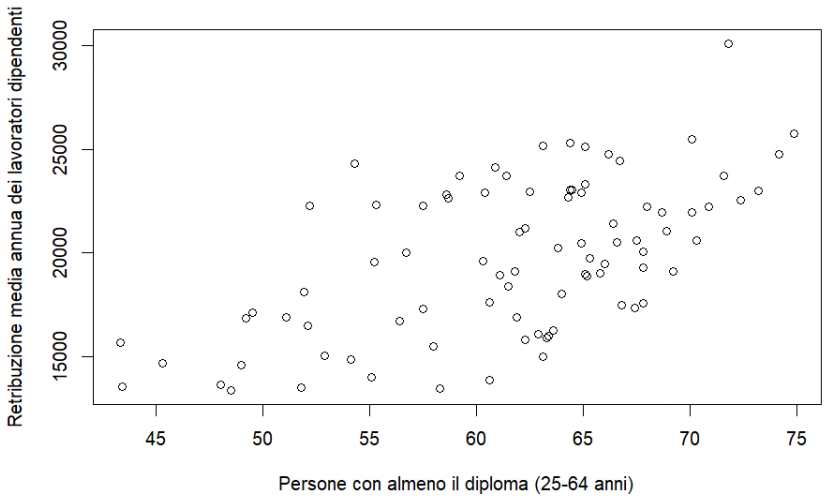
Come possiamo osservare dal boxplot riportante la distribuzione della retribuzione all'interno del nostro paese, quest'ultima è molto più corposa nelle regioni del Nord, con un picco di oltre 30.000 a Milano, e va decrescendo verso le regioni centrali e meridionali, con un picco minimo di 12.088 a Ragusa.

Persone con almeno il diploma (25-64 anni)

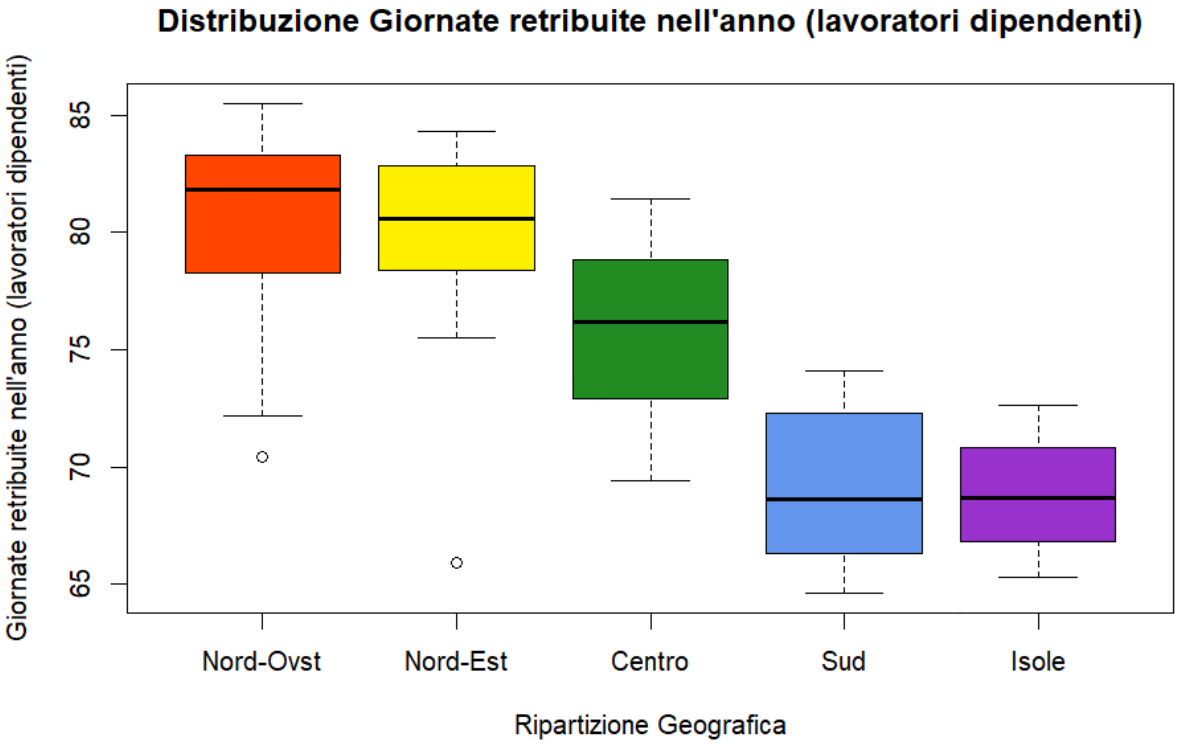


Come possiamo osservare dal boxplot riportante la distribuzione delle persone diplomate all'interno del nostro paese, si osserva come nei boxplot del Nord-ovest, est e Centro, essi siano più o meno allineati, mentre nel Sud e isole i numeri calano drasticamente, fino a toccare una media inferiore alla metà dei residenti.

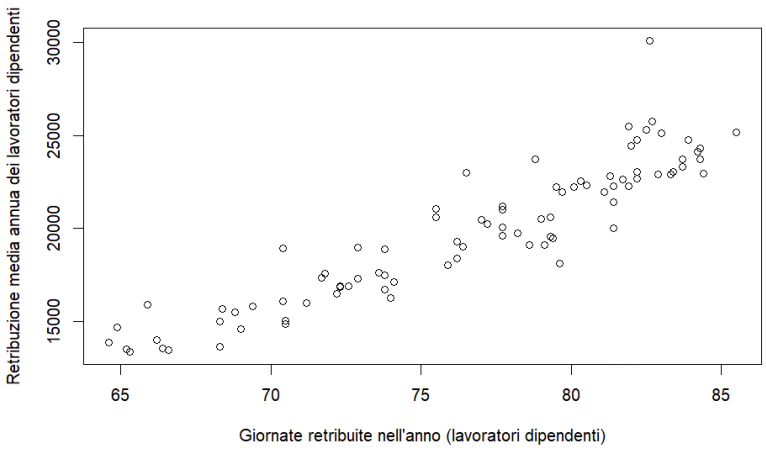
Come possiamo osservare dal plot, invece, più vi sono diplomati, più la loro retribuzione è maggiore.



Giornate retribuite nell’anno (lavoratori dipendenti)



Come possiamo osservare dal boxplot riportante la distribuzione delle giornate retribuite nell’anno, esse sono maggiori nelle regioni settentrionali, con un massimo di 85.5 a Lecco; mentre nel meridione esse decrescono fino a toccare l’infima cifra di 64.6 a Salerno.



Come possiamo osservare dal plot, invece, più aumentano le giornate retribuite annuali, più la retribuzione è maggiore.

5. Modello di regressione lineare multiplo

L'analisi della regressione multipla è una tecnica statistica che ci permette di analizzare la relazione tra una variabile dipendente e diverse variabili indipendenti (o esplicative) con l'obiettivo di prevedere quelli che potranno essere i valori assunti dalla prima a partire dalla conoscenza di quelli osservati sulle seconde. Per quanto riguarda la nostra relazione, il fenomeno attorno al quale andremo a costruire il modello sarà rappresentato dalla variabile "Retribuzione media annua dei lavoratori dipendenti", che fungerà quindi da variabile dipendente, affiancata nell'analisi da una serie di elementi esplicativi ottenuti tramite la selezione delle variabili effettuata precedentemente. Per tale procedimento utilizzeremo R.

5.1 Stima dei parametri del modello

Il primo passo per la nostra regressione multipla è omettere le righe contenenti NA tramite il comando na.omit(); successivamente passiamo alla stima dei parametri del modello attraverso il comando lm(). I principali argomenti sono la "formula" che rappresenta la descrizione simbolica del modello da stimare e "dati" che è il nome del dataframe opzionale nel quale sono presenti le variabili che figurano nel modello.

```
Call:
lm(formula = dati$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati$`Importo medio annuo pro-capite dei redditi pensionistici` +
        dati$`Patrimonio pro capite` + dati$`Reddito medio disponibile pro capite` +
        dati$`Mobilità dei laureati italiani (25-39 anni)` +
        dati$`Competenza numerica non adeguata` + dati$`Giovani che non lavorano e non studiano (Neet)` +
        dati$`Persone con almeno il diploma (25-64 anni)` + dati$`Giornate retribuite nell'anno (lavoratori dipendenti)` +
        dati$`Tasso di mancata partecipazione al lavoro` + dati$`Tasso di occupazione (20-64 anni)`,
    data = dati)

Residuals:
     Min       1Q   Median       3Q      Max
-2206.49  -676.53    -3.81   577.67  2204.57


Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                             -1.867e+04  5.577e+03  -3.348  0.00127 **
dati$`Importo medio annuo pro-capite dei redditi pensionistici`  2.235e-01  1.370e-01   1.631  0.10710
dati$`Patrimonio pro capite`                             1.216e-02  6.985e-03   1.741  0.08570 .
dati$`Reddito medio disponibile pro capite`              3.933e-01  1.441e-01   2.729  0.00788 **
dati$`Mobilità dei laureati italiani (25-39 anni)`       1.874e+00  1.517e+01   0.124  0.90203
dati$`Competenza numerica non adeguata`                 -3.578e+01  2.148e+01  -1.666  0.09990 .
dati$`Giovani che non lavorano e non studiano (Neet)`    -4.007e+01  3.923e+01  -1.021  0.31033
dati$`Persone con almeno il diploma (25-64 anni)`        1.149e+01  2.430e+01   0.473  0.63777
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`  3.541e+02  3.512e+01  10.082 1.16e-15 ***
dati$`Tasso di mancata partecipazione al lavoro`         1.064e+02  5.968e+01   1.782  0.07870 .
dati$`Tasso di occupazione (20-64 anni)`                -3.303e+01  6.159e+01  -0.536  0.59331
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 918.8 on 76 degrees of freedom
Multiple R-squared:  0.9445,    Adjusted R-squared:  0.9372
F-statistic: 129.4 on 10 and 76 DF,  p-value: < 2.2e-16
```

Tramite queste stringhe possiamo osservare i diversi dati delle variabili all'interno del nostro modello di regressione multipla:

I. ***Stima:*** l'assegnazione sulla base dei dati a nostra disposizione dei valori numerici delle variabili;

II. ***Standard error:*** la stima della deviazione standard dello stimatore, che ne esprime la sua variabilità e precisione;
III. ***t-value:*** valore osservato della statistica test;
IV. ***p-value:*** probabilità che la v.c. statistica test assuma un valore maggiore (in valore assoluto) rispetto al valore calcolato sul campione osservato. Se tale valore è inferiore al livello di significatività scelto, l'ipotesi di base verrà rifiutata. Esso è il rapporto tra la stima e lo standard error;
V. *Una serie di simboli che esprimono il livello di significatività dei coefficienti;*
VI. ***Residual standard error:*** è la stima della deviazione standard degli errori, e ci dà un'indicazione su quanto il modello di regressione si adatti bene coi nostri dati;
VII. ***Multiple R-squared e Adjusted R-squared:*** il primo è il coefficiente di determinazione multipla, che ci dà una misura della bontà di adattamento del modello: esso è 0.94, quindi il 94% del modello viene spiegato dalle variabili esplicative. Il secondo è il coefficiente corretto, che ci mostra la proporzione di variabilità del reddito medio annuo spiegata da tutte le variabili indipendenti inserite nel modello, corretta per il numero di variabili all'interno di esso. Il coefficiente R^2 corretto è pari a 0.93: ciò significa che la nostra regressione spiega circa il 93% della varianza della variabile dipendente;
VIII. ***F-statistic:*** ci dice se il modello debba essere scartato nella sua interezza oppure possa essere ritenuto valido; dato che l'F statistico è maggiore dell'F tabulato possiamo rifiutare l'ipotesi nulla che il modello sia da non considerare nella sua totalità.

5.2 Intervalli di confidenza dei regressori

Ora passiamo a calcolarci gli intervalli di confidenza del modello tramite il comando confint(). Si noti che R per default ci fornirà gli intervalli di confidenza pari al 95%

	2.5 %	97.5 %
(Intercept)	-2.977866e+04	-7.562514e+03
dati$`Importo medio annuo pro-capite dei redditi pensionistici`	-4.947527e-02	4.964273e-01
dati$`Patrimonio pro capite`	-1.749940e-03	2.607512e-02
dati$`Reddito medio disponibile pro capite`	1.063073e-01	6.802657e-01
dati$`Mobilità dei laureati italiani (25-39 anni)`	-2.834818e+01	3.209663e+01
dati$`Competenza numerica non adeguata`	-7.855334e+01	7.002574e+00
dati$`Giovani che non lavorano e non studiano (Neet)`	-1.182102e+02	3.806858e+01
dati$`Persone con almeno il diploma (25-64 anni)`	-3.690535e+01	5.987647e+01
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`	2.841301e+02	4.240208e+02
dati$`Tasso di mancata partecipazione al lavoro`	-1.249747e+01	2.252355e+02
dati$`Tasso di occupazione (20-64 anni)`	-1.557053e+02	8.963944e+01

Questa schermata permette di farci comprendere la proprietà fondamentale degli intervalli di confidenza: essi permettono di confermare l'ipotesi di non significatività delle variabili esplicative. Di fatti, non deve mai accadere che l'intervallo di

confidenza possa contenere lo zero e, se ciò accade, significa che la variabile considerata non sarà significativa.

5.3 Verifica della multicollinearità

A questo punto decidiamo di effettuare la verifica di multicollinearità: anche se abbiamo già fatto attenzione nella fase di selezione dei dati a questo possibile fenomeno, che si sviluppa quando due o più variabili indipendenti di un modello di regressione sono altamente correlate tra loro. Tramite la funzione vif() (Variance Inflation Factor), che determina la forza della correlazione tra le variabili indipendenti, elimineremo tutte quelle variabili che risulteranno avere un valore significativamente superiore a 5, punto in cui la multicollinearità è probabile.

```
dati$`Importo medio annuo pro-capite dei redditi pensionistici`
                                                        6.182504
                                    dati$`Patrimonio pro capite`
                                                       10.151119
                        dati$`Reddito medio disponibile pro capite`
                                                       26.422768
                dati$`Mobilità dei laureati italiani (25-39 anni)`
                                                        5.162191
                           dati$`Competenza numerica non adeguata`
                                                        6.540056
             dati$`Giovani che non lavorano e non studiano (Neet)`
                                                        9.963337
                 dati$`Persone con almeno il diploma (25-64 anni)`
                                                        3.119174
    dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`
                                                        4.240366
                  dati$`Tasso di mancata partecipazione al lavoro`
                                                       38.122104
                           dati$`Tasso di occupazione (20-64 anni)`
                                                       40.765191
```

Avendo osservato la possibilità di presenza di multicollinearità per molte variabili, abbiamo deciso di tenere nel nostro modello di regressione solo le variabili “Persone con almeno il diploma (25-64 anni)” e “Giornate retribuite nell’anno (lavoratori dipendenti)”; otteniamo il nostro modello significativo:

```
Call:
lm(formula = dati$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati$`Persone con almeno il diploma (25-64 anni)` + dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`,
    data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-2501.3  -860.9   -99.9   631.5  6130.4

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        -26217.35    1914.33 -13.695  < 2e-16 ***
dati$`Persone con almeno il diploma (25-64 anni)`      98.05      22.49   4.361 3.66e-05 ***
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`    522.26      27.88  18.735  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1319 on 84 degrees of freedom
Multiple R-squared:  0.8736,    Adjusted R-squared:  0.8706
F-statistic: 290.3 on 2 and 84 DF,  p-value: < 2.2e-16
```

5.4 Scelta dei regressori significativi

Uno dei problemi che spesso lo statistico deve affrontare quando effettua l'analisi della regressione è quello della scelta dei regressori da inserire nel modello per definire la variabile oggetto del lavoro. Questo è un problema complicato in quanto bisognerebbe includere nel modello solo quelle variabili esplicative la cui variazione apporta un contributo reale alla variazione della variabile dipendente. Una possibilità è quella di considerare un approccio che permetta di individuare quali variabili esplicative consentono effettivamente di costruire un modello adeguato senza dover ricorrere all'uso di tutte le variabili considerate. Una volta fatto ciò si andrà ad effettuare una analisi dei residui per valutarne l'adeguatezza. Tra gli algoritmi di scelta delle variabili, il problema della regola di arresto viene risolto con riferimento all'indice AIC (Akaike Information Criterion), abbiamo:

1) ***Backward elimination:*** si parte considerando il modello che include tutte le variabili a disposizione. Si fissa un livello di significatività. La variabile con il coefficiente di regressione meno significativo in base al test t viene eliminata; quindi, si calcolano di nuovo le stime e si ripete il procedimento fino a che non ci sono più covariate che risultano non significative al livello prefissato.
2) ***Forward selection:*** si parte con una sola covariata, quella con la maggiore correlazione significativa (test t) con la variabile risposta. Si fissa un livello di significatività. La seconda variabile da inserire è quella che presenta il coefficiente di correlazione parziale più elevato e significativo, si prosegue si inserendo una successiva variabile dipendente. Il procedimento ha fine quando il coefficiente di correlazione parziale dell'ultima variabile inserita non è più significativa rispetto al livello prefissato.
3) ***Stepwise regression:*** è una combinazione dei due criteri precedenti. La selezione delle covariate da includere nel modello avviene come nella forward selection. Aggiungendo successivamente una nuova variabile, i coefficienti di

regressione delle variabili già incluse potrebbero risultare singolarmente non significativi a causa della forte correlazione con la nuova variabile. Perciò dopo l'inserimento di ciascuna variabile il modello viene riconsiderato per verificare se vi è qualche variabile da eliminare (come nella backward elimination). Applichiamo, la prima e la terza procedura in modo tale da avere una conferma sul modello, andando a fissare un livello di significatività pari a 0.05.

```
Start:  AIC=1253.12
dati$`Retribuzione media annua dei lavoratori dipendenti` ~ dati$`Persone con almeno il diploma (25-64 anni)` +
    dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`

                                                        Df Sum of Sq       RSS    AIC
<none>                                                              146223133 1253.1
- dati$`Persone con almeno il diploma (25-64 anni)`      1  33099145 179322278 1268.9
- dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`  1 610999917 757223049 1394.2

Call:
lm(formula = dati$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati$`Persone con almeno il diploma (25-64 anni)` + dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`,
    data = dati)

Coefficients:
                                          (Intercept)
                                            -26217.35
          dati$`Persone con almeno il diploma (25-64 anni)`
                                                98.05
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`
                                               522.26
```

```
Start:  AIC=1253.12
dati$`Retribuzione media annua dei lavoratori dipendenti` ~ dati$`Persone con almeno il diploma (25-64 anni)` +
    dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`

                                                        Df Sum of Sq       RSS    AIC
<none>                                                              146223133 1253.1
- dati$`Persone con almeno il diploma (25-64 anni)`      1  33099145 179322278 1268.9
- dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`  1 610999917 757223049 1394.2

Call:
lm(formula = dati$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati$`Persone con almeno il diploma (25-64 anni)` + dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`,
    data = dati)

Coefficients:
                                          (Intercept)
                                            -26217.35
          dati$`Persone con almeno il diploma (25-64 anni)`
                                                98.05
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`
                                               522.26
```

Osserviamo che in entrambi i casi le variabili rimanenti all'interno della regressione multipla sono le stesse; quindi, le due procedure diverse hanno stimato lo stesso modello. Ora verifichiamo la significatività delle componenti del modello.

```
Call:
lm(formula = dati$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati$`Persone con almeno il diploma (25-64 anni)` + dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`,
    data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-2501.3  -860.9   -99.9   631.5  6130.4

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        -26217.35    1914.33 -13.695  < 2e-16 ***
dati$`Persone con almeno il diploma (25-64 anni)`      98.05      22.49   4.361 3.66e-05 ***
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`    522.26      27.88  18.735  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1319 on 84 degrees of freedom
Multiple R-squared:  0.8736,    Adjusted R-squared:  0.8706
F-statistic: 290.3 on 2 and 84 DF,  p-value: < 2.2e-16
```

A questo punto il modello è significativo secondo tutte le sue componenti, i due indici R^2 sono molto simili tra loro e l'F statistico afferma la validità del nostro lavoro.

5.5 ANOVA

Ora andiamo a verificare quanto appena fatto tramite un test ANOVA. Esso è utile per confrontare due o più modelli di regressione che differiscono per il numero di variabili esplicative inserite; si usa il comando anova()che mette in evidenza se le variabili in più o in meno di un modello rispetto all'altro apportano oppure no un contributo significativo nello spiegare la variabile risposta.

```
Call:
lm(formula = dati$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`,
    data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-3472.5  -755.4  -251.2   553.4  6775.4

Coefficients:
                                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       -24618.53    2068.42  -11.90   <2e-16 ***
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`    580.32      26.96   21.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1452 on 85 degrees of freedom
Multiple R-squared:  0.845,     Adjusted R-squared:  0.8432
F-statistic: 463.3 on 1 and 85 DF,  p-value: < 2.2e-16

Call:
lm(formula = dati$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati$`Persone con almeno il diploma (25-64 anni)` + dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`,
    data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-2501.3  -860.9   -99.9   631.5  6130.4

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        -26217.35    1914.33 -13.695  < 2e-16 ***
dati$`Persone con almeno il diploma (25-64 anni)`      98.05      22.49   4.361 3.66e-05 ***
dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`    522.26      27.88  18.735  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1319 on 84 degrees of freedom
Multiple R-squared:  0.8736,    Adjusted R-squared:  0.8706
F-statistic: 290.3 on 2 and 84 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Model 1: dati$`Retribuzione media annua dei lavoratori dipendenti` ~ dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`
Model 2: dati$`Retribuzione media annua dei lavoratori dipendenti` ~ dati$`Persone con almeno il diploma (25-64 anni)` +
    dati$`Giornate retribuite nell'anno (lavoratori dipendenti)`
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1     85 179322278
2     84 146223133  1  33099145 19.014 3.66e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Essendo il p-value minore del livello di significatività, ovvero 0.05, siamo portati a rifiutare l'ipotesi nulla secondi cui l'apporto della variabile "Persone con almeno il diploma (25-64 anni)" non contribuisca significativamente alla spiegazione della variabile dipendente "Retribuzione media annua dei lavoratori dipendenti".

5.7 Commento dei risultati

Ottenuto il modello di regressione significativo in tutte le sue componenti, ci apprestiamo a commentare brevemente i segni delle variabili esplicative per comprendere in che modo vadano ad illustrare la nostra variabile dipendente:

- "Persone con almeno il diploma(25-64 anni)": un aumento unitario delle persone con il diploma comporta una variazione positiva della retribuzione media annua dei lavoratori dipendenti pari a 98.05.
- "Giornate retribuite nell'anno (lavoratori dipendenti)": un aumento unitario delle giornate retribuite all'interno dell'anno comporta una variazione positiva della retribuzione media dei lavoratori dipendenti pari a 522.26.

6. Test di specificazione e diagnostica

L'ultimo passo fondamentale è quello di verificare la veridicità delle ipotesi base del modello stesso, sia con i test che attraverso una analisi diagnostica tramite dei grafici.

6.1 Test t di student

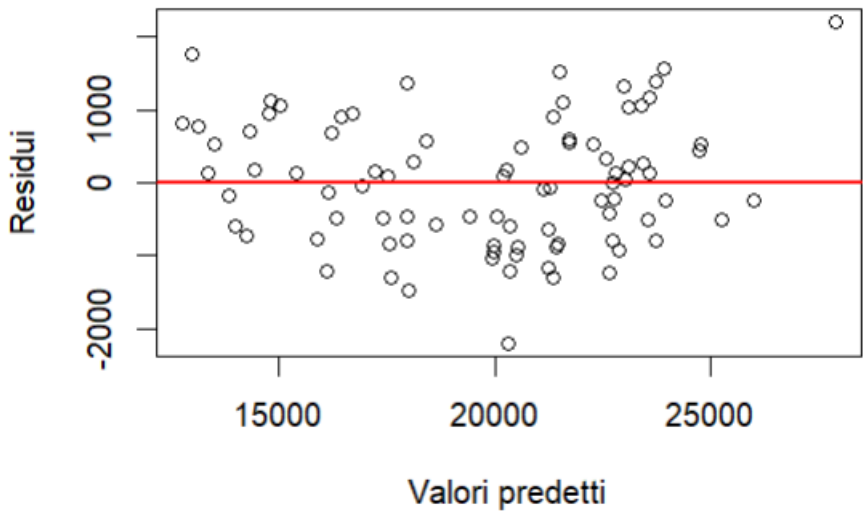
Il t-test occorre per verificare che la media dei residui non sia significativamente diversa da 0.

```
        One Sample t-test

data:  residui
t = 5.2731e-16, df = 86, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -184.087  184.087
sample estimates:
   mean of x
4.882971e-14
```

Abbiamo creato un vettore dei residui su cui abbiamo applicato il test. Siccome il p-value è 1, accettiamo l'ipotesi di base secondo cui la media degli errori non è significativamente diversa da 0.

Per verificare la linearità occorre tracciate il grafico dei residui (ordinata) verso i valori previsti (ascissa). I punti dovrebbero essere distribuiti in modo simmetrico intorno ad una linea orizzontale con intercetta uguale a zero. Andamenti di tipo diverso indicano la presenza di non linearità.



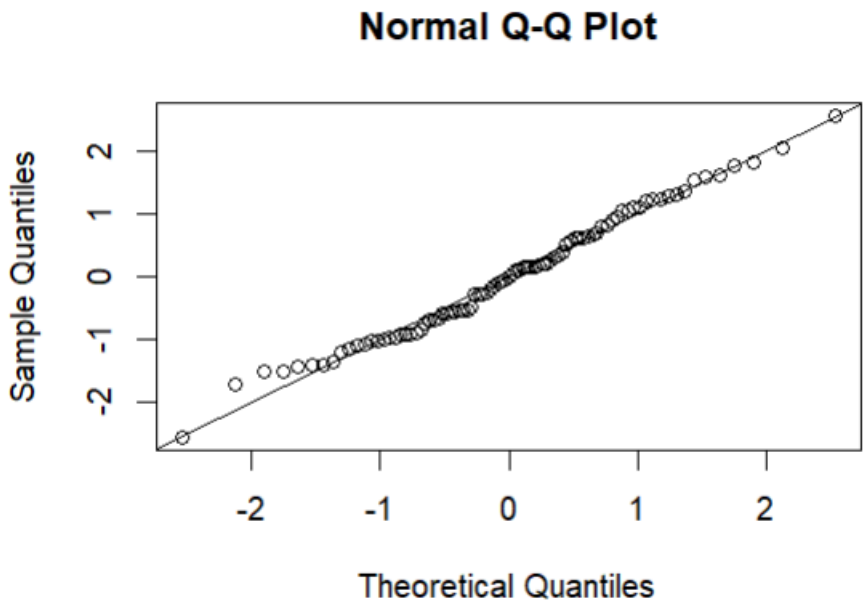
6.2 Test di Shapiro-Wilks

Il test di Shapiro-WIlks verifica la normalità della distribuzione degli errori.

```
        Shapiro-Wilk normality test

data:  residui
W = 0.98813, p-value = 0.6175
```

Notiamo che il p-value è maggiore del livello di significatività; perciò, accettiamo l'ipotesi nulla che ci restituisce la normalità della distribuzione dei residui.



Affinché l'ipotesi di normalità sia confermata tecnicamente al 100% tutti i punti dovrebbero giacere perfettamente sulla retta e ciò non accade. Ciò è dovuto alla presenza di alcuni dati che si discostano dalla normalità e per i quali la retta di regressione non riesce a interpolare correttamente; il risultato vede comunque la maggior parte dei punti ricondursi alla diagonale perciò rimane statisticamente corretto.

6.3 Test di Breusch-Pagan

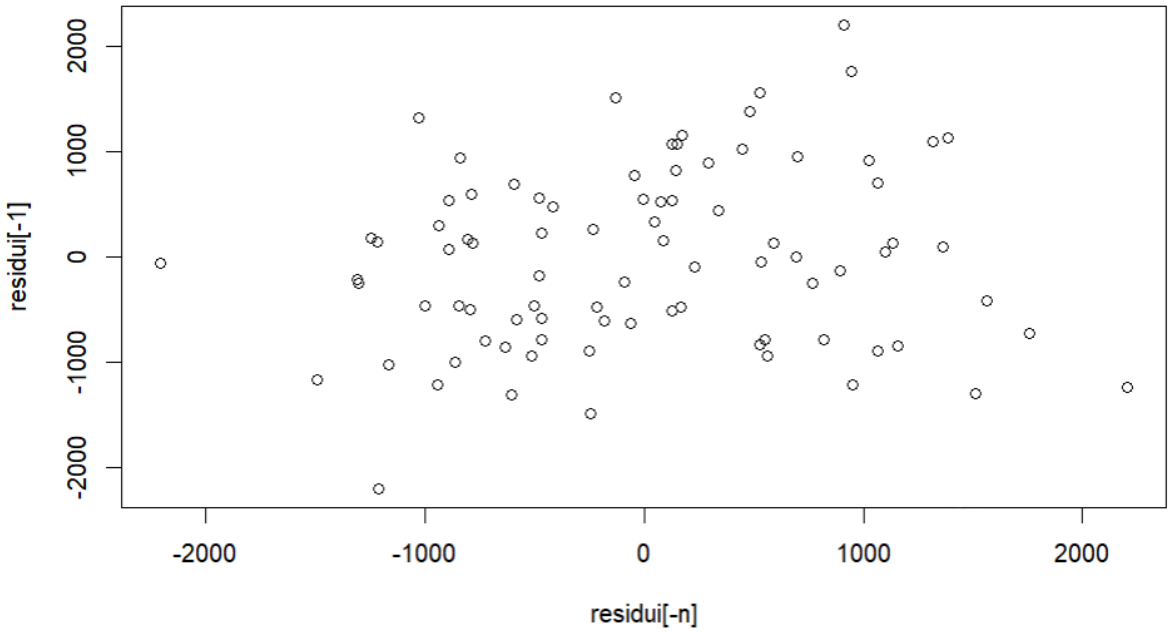
Il test Breusch-Pagam verifica l'omoschedasticità dei residui, ovvero che la varianza dei residui sia la stessa per ogni i:

```
        studentized Breusch-Pagan test

data:  modello_sign
BP = 2.2155, df = 2, p-value = 0.3303
```

p-value maggiore del livello di significatività, perciò, accettiamo l'ipotesi di base di omoschedasticità dei residui al 5%.

Da un punto di vista grafico occorrerà tracciare il grafico dei residui in valore assoluto (ordinata) rispetto ai valori stimati col modello.



La dispersione verticale dei residui si prospetta più o meno costante; perciò, siamo portati ad accettare l'ipotesi di omoschedasticità dei residui.

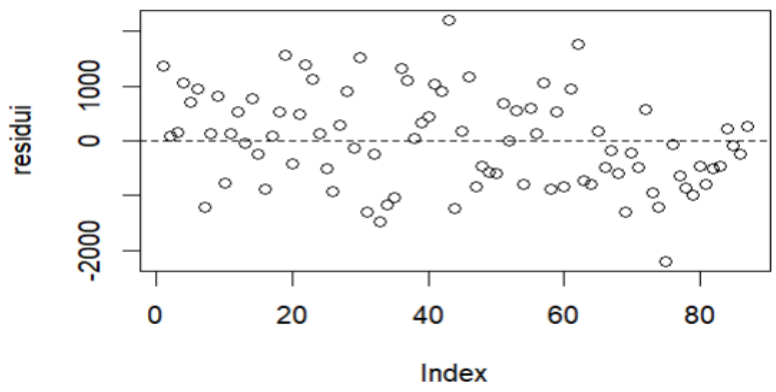
6.4 Test di Durbin-Watson:

Il test Durbin-Watson serve per affermare l'assenza di correlazione seriale tra i residui.

```
        Durbin-Watson test

data:  modello_sign
DW = 1.9721, p-value = 0.4158
alternative hypothesis: true autocorrelation is greater than 0
```

Come vediamo, anche qui il p-value è maggiore di 0.05 perciò possiamo accettare l'ipotesi nulla di assenza di correlazione seriale tra i residui. Tutto ciò è confermato dal fatto che la statistica DW assume valori prossimi a 2.



7. Modello di regressione logistico

Per la parte finale della nostra analisi, creiamo un modello di regressione logit che ci permetta di visualizzare e di prevedere, sulla base dei nostri dati, quante delle province analizzate presenteranno un livello alto di "retribuzione media annua dei lavoratori dipendenti" e quante invece un livello basso. Per realizzare il modello dobbiamo innanzitutto rendere la nostra variabile dipendente, originariamente numerica, una variabile binaria, trattando il valore 1 come "alta retribuzione media annua" e 0 come "bassa retribuzione media annua". Abbiamo quindi provveduto a trasformare la nostra variabile numerica, in una variabile dicotomica che assumesse valori pari a 1 per una soglia superiore ad un valore pari a 20.000, e valori pari a 0 per valori inferiori a 20.000. Occorre specificare, inoltre, che così come per il modello di regressione multiplo precedentemente svolto, abbiamo eliminato tutti i casi di dati NA.

Partiamo col modello logit presentando le variabili esplicative rimaste alla fine del modello di regressione multipla significativo. Successivamente stimiamo il modello tramite il comando glm() di R:

```
Call:
glm(formula = dati_l$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati_l$`Persone con almeno il diploma (25-64 anni)` + dati_l$`Giornate retribuite
nell'anno (lavoratori dipendenti)`,
    family = binomial(link = "logit"), data = dati_l)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.05560  -0.14757  -0.00318   0.28736   2.07465

Coefficients:
                                                              Estimate Std. Error
(Intercept)                                                 -64.52839   15.77343
dati_l$`Persone con almeno il diploma (25-64 anni)`           0.21864    0.08255
dati_l$`Giornate retribuite nell'anno (lavoratori dipendenti)`   0.65372    0.16055
                                                            z value Pr(>|z|)
(Intercept)                                                  -4.091 4.30e-05 ***
dati_l$`Persone con almeno il diploma (25-64 anni)`           2.649  0.00808 **
dati_l$`Giornate retribuite nell'anno (lavoratori dipendenti)`   4.072 4.67e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.596  on 86  degrees of freedom
Residual deviance:  43.254  on 84  degrees of freedom
AIC: 49.254

Number of Fisher Scoring iterations: 7
```

Ora passiamo a verificare la significatività delle componenti del nostro modello:

```
Call:
glm(formula = dati_l$`Retribuzione media annua dei lavoratori dipendenti` ~
    dati_l$`Mobilità dei laureati italiani (25-39 anni)` + dati_l$`Giornate retribuite
nell'anno (lavoratori dipendenti)`,
    family = binomial(link = "logit"), data = dati_l)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.17989  -0.09210  -0.00004   0.28701   2.08032

Coefficients:
                                                                 Estimate
(Intercept)                                                     -42.66269
dati_l$`Mobilità dei laureati italiani (25-39 anni)`              0.27555
dati_l$`Giornate retribuite nell'anno (lavoratori dipendenti)`    0.56909
                                                                Std. Error z value
(Intercept)                                                       13.35619  -3.194
dati_l$`Mobilità dei laureati italiani (25-39 anni)`               0.08379   3.288
dati_l$`Giornate retribuite nell'anno (lavoratori dipendenti)`     0.17177   3.313
                                                                Pr(>|z|)
(Intercept)                                                     0.001402 **
dati_l$`Mobilità dei laureati italiani (25-39 anni)`            0.001007 **
dati_l$`Giornate retribuite nell'anno (lavoratori dipendenti)`  0.000923 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.596  on 86  degrees of freedom
Residual deviance:  32.813  on 84  degrees of freedom
AIC: 38.813

Number of Fisher Scoring iterations: 8
```

Osserviamo che tutte le variabili presenti all’interno del nostro modello logit hanno un p-value molto basso, perciò esse sono tutte significative.

7.1 Predict

Fatto ciò, procediamo a tradurre l’esito appena raggiunto in termini di probabilità grazie al comando predict() e andiamo a costruire la nostra tabella di conclusione fissando una soglia di 0.25 oltre la quale il fenomeno verrà osservato sotto una caratterizzazione “forte”.

```
          1           2           3           4           5           6           7           8           9          10          11          12
-5.36041129 -5.08131323 -3.56757225 -7.17313218 -11.81262214 -7.99735839 -8.90731521 -10.35520743 -20.99513322 -15.10724816 -9.35121328 -11.51972100
         13          14          15          16          17          18          19          20          21          22          23          24
-7.91051381 -10.77694480 14.98202997  3.12184678  2.31793219  6.24341696  7.41750390  2.71750886  1.43315000  5.64619224 -4.41585912  3.50227671
         25          26          27          28          29          30          31          32          33          34          35          36
 3.56516242  2.38806153 -2.68755131 -4.48287275 -6.05653637  0.84491622 -2.72711145  4.41334751 -5.54255985  2.54736519 -2.04182639  3.87847694
         37          38          39          40          41          42          43          44          45          46          47          48
 4.43793971  4.06115705  2.28265263  6.49024941  2.66424742  3.52322165 12.41759284  2.99967176 -1.32295819  5.60726542  2.19036259 -2.62048680
         49          50          51          52          53          54          55          56          57          58          59          60
-3.35427206  0.13149347 -7.35940866  1.66384965  0.03269489  0.93186258  1.60572519  6.98143192  5.32510252 -0.86948675 -0.50164414 -4.68713111
         61          62          63          64          65          66          67          68          69          70          71          72
-11.61793715 -17.46747782 -12.28169484 -7.54748333 -14.58311583 -8.48377965 -13.16282976 -14.01587787  1.56679837  5.15675703 -3.74670687 -4.59309378
         73          74          75          76          77          78          79          80          81          82          83          84
 0.29201288  1.96632416  2.27842293  0.48072087  3.03135494  0.35912828 -0.64603074 -5.38204044  3.54776297  4.44095366 -1.36426563  4.63923503
         85          86          87
 1.30737860  3.00385057  3.21715076
```

Ora passando alla tabella di classificazione osserviamo che:

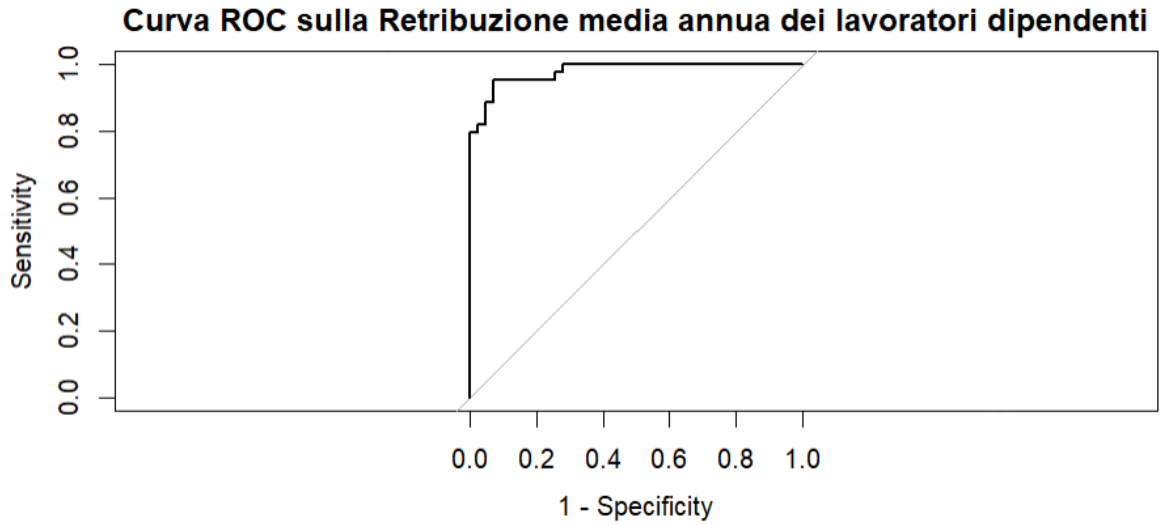
```
    0  1
0  44  0
1   0 43
```

-44: numero di province previste con una retribuzione media annua dei lavoratori dipendenti alta e che effettivamente hanno confermato tale risultato;

-43: numero di province con una retribuzione media annua dei lavoratori dipendenti bassa e che effettivamente hanno confermato tale previsione;

7.2 Curva ROC

Questi risultati ottenuti dipendono dalla soglia che è stata scelta a priori e per poter avere a disposizione un quadro generale di come è il nostro modello di regressione logit e come questo classificatore viene considerato al variare di diverse soglie possiamo fare ricorso alla cosiddetta curva "ROC" (Receiver Operating Characteristic), che confronta il grado di sensitività (frazione di veri positivi) con il termine 1-specificità (frazione di falsi positivi). Il risultato che otteniamo sarà il seguente:



8. Conclusione

Siamo ormai giunti alla conclusione delle nostre analisi e sembra dunque doveroso presentare una breve conclusione sul lavoro svolto e sui risultati a cui siamo giunti. Ricapitolando le varie fasi di analisi, i dati presentati durante l'esame descrittivo del dataset hanno portato alla luce la vera essenza della retribuzione media annua e delle variabili ad essa più correlate. La retribuzione è risultata particolarmente diffusa maggiormente nella parte settentrionale italiana. In prima istanza abbiamo potuto verificare la relazione tra questa variabile di interesse e le singole variabili che, durante la selezione dei dati, sono risultati più influenti nei confronti della variabile dipendente. Siamo così passati ad un modello di regressione lineare multiplo in cui l'obiettivo è stato proprio quello di analizzare nel complesso la relazione tra la variabile dipendente con le altre variabili esplicative. Il risultato a cui siamo giunti è stata l'intrinseca dipendenza della retribuzione da variabili quali le persone diplomate e le giornate retribuite l'anno (variabili che, durante l'analisi descrittiva presentavano valori favorevoli prevalentemente al nord e valori sfavorevoli nel sud e isole). In ultima istanza abbiamo deciso di trasformare la retribuzione media annua in una variabile qualitativa dicotomica con l'obiettivo di svolgere una regressione logistica sul modello. Abbiamo affidato infatti il valore 1 a quelle osservazioni della retribuzione che superavano una soglia di 20.000, tale da poter essere considerate come maggiormente contribuenti al fenomeno, e con 0 quelle osservazioni con valori al di sotto della soglia e quindi meno contribuenti della retribuzione. Effettuando una valutazione sulle previsioni, abbiamo tracciato la curva ROC, la quale mostra il trade-off tra sensibilità e specificità.

A cura di:

- Felice Attanasio
- Giorgio Cantalamessa
- Alessandro Alivernini