A quantitative analysis of an economic database.
Quantitative Methods for Economics 1st intermediate project

Quantitative Report

Regression analysis

Stefano Blando

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Dipartimento di Economia e Finanza

FACOLTA' DI ECONOMIA

Analisi Statistica e di Regressione per;

Salario Netto (RETRIC) in Italia (da LFS Istat per 2015-3rd quarter)

Analista:

BLANDO Stefano | 0334525

Supervisore:

Professoressa Chiara Perricone

2022/2023

Contents

Introduzione

Il report di riferimento presenta un'analisi statistica delle retribuzioni mensili nette in Italia attraverso l'utilizzo dei dati del Labour Force Survey del 2015 dell'Istituto Nazionale di Statistica (ISTAT).

L'obiettivo dell'analisi è quello di valutare le distribuzioni deli salari in base al livello di istruzione, alla regione di residenza, all'età, al genere, alla condizione occupazionale e alla nazionalità dei lavoratori.

Per raggiungere tale obiettivo il report descrive i dati utilizzati e le metodologie adoperate per l'analisi statistica, tra cui la definizione di specifici indicatori socio-demografici e la stima di modelli di regressione lineare multivariata.

L'analisi evidenzia differenze significative nella distribuzione dei salari netti mensili in Italia in base a molteplici variabili, come ad esempio il livello di istruzione, la regione di residenza, l'età e al genere.

I lavoratori con un livello di istruzione più elevato tendono a guadagnare di più, mentre le regioni del sud tendono a presentare salari netti mensili inferiori rispetto alle regioni del nord. Si osservano anche disparità salariali tra uomini e donne, con quest'ultime che in media guadagnano meno degli uomini.

Anche l'età risulta essere un fattore discriminante nella distribuzione del salario netto mensile, con i lavoratori più anziani che tendono a guadagnare di più rispetto ai lavoratori più giovani.

Il report conclude con una sintesi dei principali risultati dell'analisi statistica e le relative osservazioni sulla capacità esplicativa dei modelli repressivi elaborati.

L'analisi offre una visione completa del mercato del lavoro italiano e fornisce informazioni utili per la comprensione e la valutazione della situazione economica del paese. .

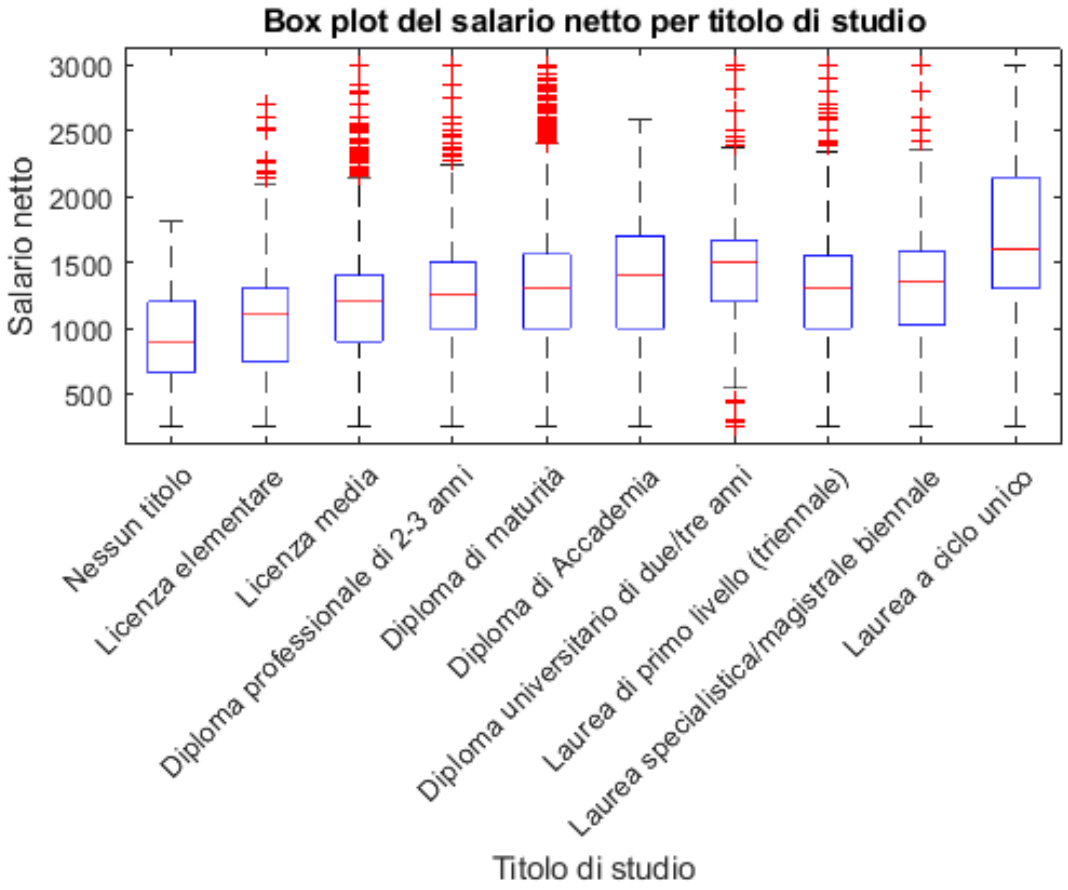
Analisi Statistica

Il punto di partenza dell'analisi è stato quello di fornire alcune statistiche descrittive e una rappresentazione grafica del salario mensile netto, preso in modo incondizionato e successivamente anche condizionato rispetto a due variabili qualitative scelte.

Statistica Descrittiva

In particolare, al fine di indagare sulla distribuzione dei salari netti mensili in Italia, sono state utilizzate le variabili "TISTUD", che indica il livello di istruzione e "REG", che considera la regione di appartenenza,

In riferimento alla variabile "TISTUD" è stato elaborato un box plot (figura 1) della distribuzione del salario netto mensile diviso per titolo di studio, dal quale si evince una relazione tra un più alto grado di formazione ed un maggiore livello di salario netto mensile medio, nonostante la presenza di *outliers* nella distribuzione delle singole modalità.

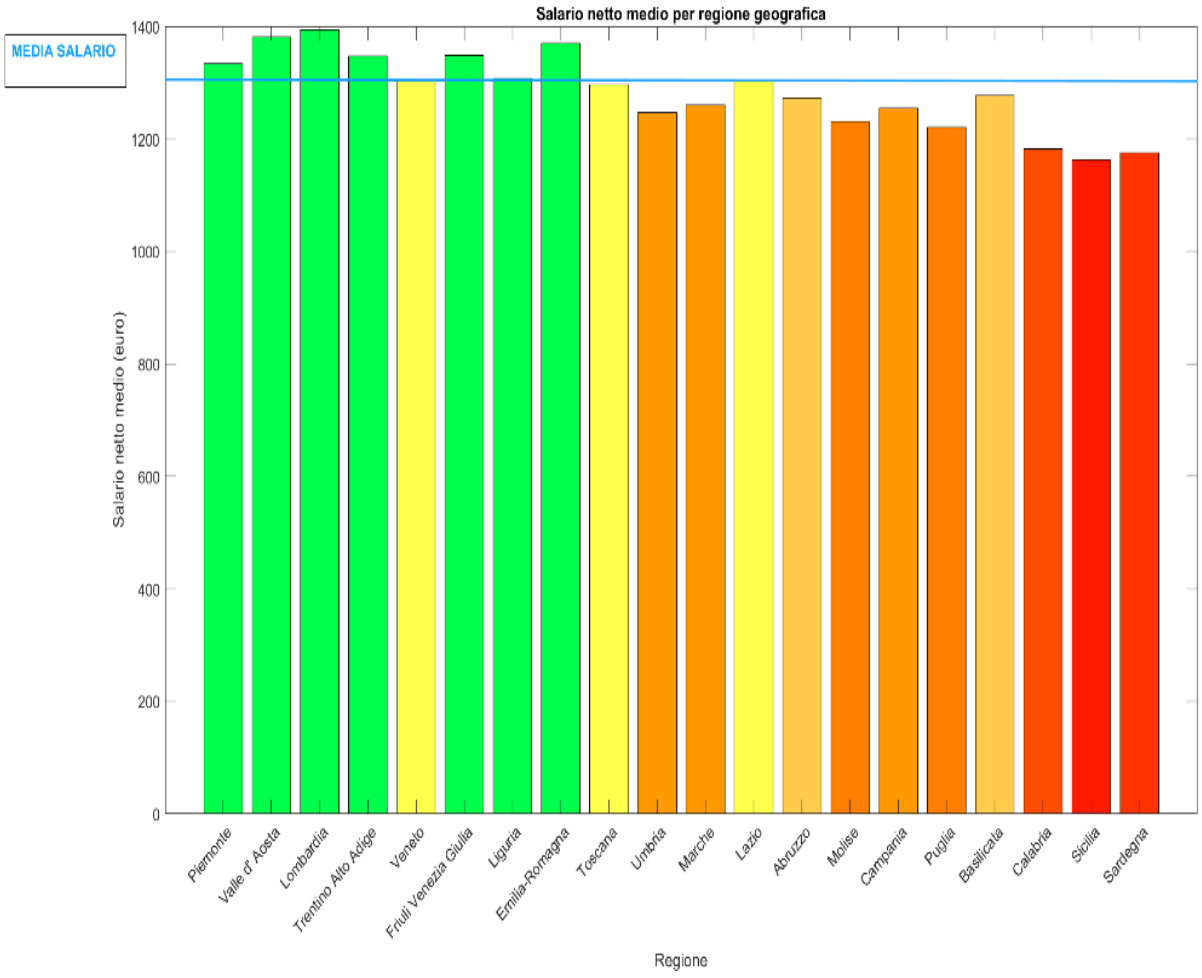


(Figura 1)

Per ciò che concerne la Regione d'appartenenza è stato creato un grafico a barre della distribuzione del salario netto medio mensile suddiviso per area geografica.

Come è possibile osservare nella figura 2, i salari medi variano notevolmente tra le diverse regioni, con le regioni del Nord che presentano una distribuzione dei salari netti mensili mediamente superiore rispetto alle regioni del Sud. In particolare, le regioni che mostrano i salari netti mensili più elevati sono la Valle d'Aosta, la Lombardia e l'Emilia-Romagna, mentre quelle con i salari netti mensili più bassi sono la Calabria, la Sicilia e la Sardegna.

In sintesi, l'analisi descrittiva condotta evidenzia che il salario netto mensile in Italia varia in base alla regione di residenza e al grado di formazione, e che la distribuzione del salario netto mensile nel campione non segue una distribuzione normale.



(Figura 2)

Significatività statistica

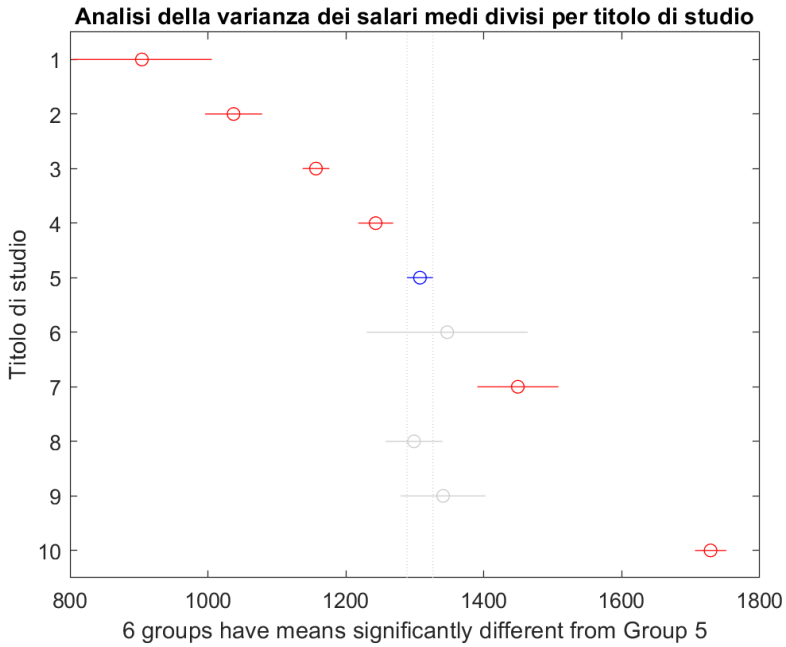
Al fine di confermare questi primi risultati ottenuti dall'analisi della variabile "TISTUD" si è deciso di analizzare la variabile in due specifici sottogruppi:

- subgroup_1edu: modalità 1-5 = individui senza un titolo di laurea
- subgroup_2edu: modalità 7-10 = individui con almeno un titolo di laurea

Dall'analisi della varianza (figura 9) si denota la presenza di una differenza statisticamente significativa tra le medie del salario mensile dei due sottogruppi.

In particolare, si nota che esiste una differenza statisticamente significativa tra i salari di chi ha un diploma di maturità (modalità 5) e tutti gli individui con titoli minori, ma anche chi ha un diploma universitario (modalità 7) o un titolo di laurea a ciclo unico (modalità 10); c'è una differenza significativa tra chi ha un diploma di Accademia (modalità 6) e chi ha fino alla licenza media (modalità 1-3) o un titolo di laurea a ciclo unico (modalità 10); tra chi ha una laurea triennale (modalità 8) e chi ha un diploma universitario (modalità 7) o una laurea a ciclo unico (modalità 10); tra chi ha un diploma universitario (modalità 7) e chi ha una laurea a ciclo unico (modalità 10) oppure un titolo fino al diploma di maturità (modalità 1-5).

Nonostante le persone con titolo di studio più elevato abbiano salari netti medi mensili tendenzialmente più elevati, esistono differenze significative tra i salari medi di diversi titoli di studio superiori: nello specifico, conseguire una laurea specialistica a ciclo unico (modalità 10) consente di ottenere una salario mediamente più alto di ogni altro titolo, ma tra i salari medi di chi ha un diploma di maturità (modalità 5) e titoli di laurea triennale o specialistica (modalità 8 e 9) non risultano esserci differenze statisticamente significative.



(Figura 3)

Percentili

Rispetto alla variabile "TISTUD" è possibile notare che la maggior parte degli individui appartenenti al quinto percentile inferiore della distribuzione del salario medio netto mensile sono quelli con licenza media, mentre la maggior parte degli individui appartenenti al quinto percentile superiore sono quelli in possesso del diploma di maturità: ciò è dovuto alla più elevata frequenza relativa di rispondenti (0.29) con licenzia media per il sottogruppo delle modalità di TISTUD con salario medio inferiore alla media ("Senza titolo", " Licenza elementare", "Licenza media", "Diploma professionale", "Laurea di primo livello"), e la maggiore frequenza relativa di rispondenti (0.40) con diploma di maturità per il sottogruppo delle modalità di TISTUD complementare al primo con salario medio superiore alla media ("Diploma di Accademia", "Diploma universitario di 2/3 anni", "Laurea specialistica", "Laurea specialistica a ciclo unico"), rispetto al totale degli intervistati.

Analisi di Regressione

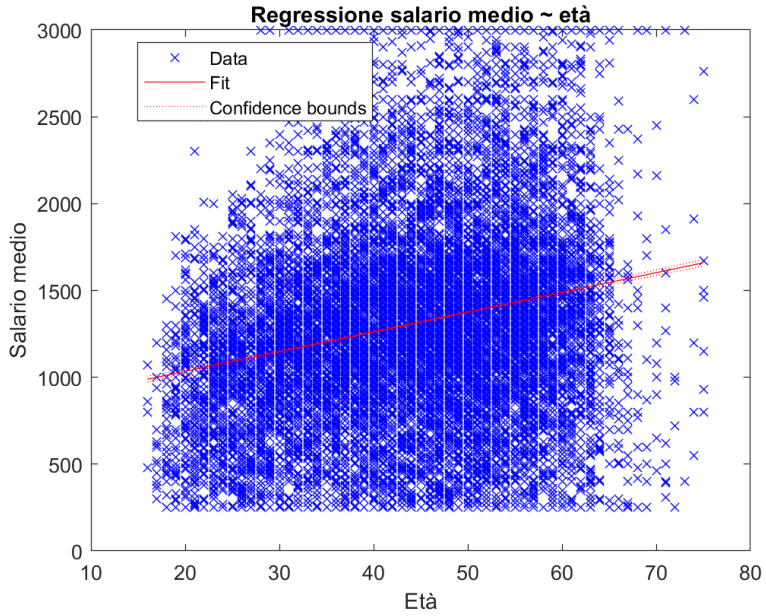
La nostra analisi statistica di regressione si concentra sul salario netto mensile in Italia. Attraverso questa analisi, abbiamo esplorato le variabili che influenzano il salario netto, identificando fattori chiave come l'età, l'istruzione, l'esperienza lavorativa e altri indicatori economici. Questo ci ha permesso di comprendere meglio i determinanti del salario netto in Italia e fornire una panoramica dettagliata per un report quantitativo informativo.

Regressione Univariata

Un primo modello esplorativo di regressione tra salario medio ("RETRIC") ed età ("ETAM") evidenzia una trascurabile correlazione tra le due variabili quantitative, presentando un valore dell'R^2 adjusted pari a 0,0564.

Infatti, è anche possibile notare un valore del RMSE molto alto, pari a 508, sintomo di un significativo errore medio commesso dalla regressione di riferimento. Dall'analisi della variabilità è inoltre possibile notare una distribuzione dei residui non omogenea: il valore riportato dal test dei coefficienti risulta essere pari a 26.8983.

Il modello presenta, quindi, eteroschedasticità.



(Figura 4)

Omitted variable bias

Dati i non incoraggianti risultati del primo modello regressivo è necessario considerare il problema di "*omitted variable bias*": analizzando la relazione tra le variabili di riferimento individuate si trascura l'influenza di un'ulteriore variabile "Z" collegata con la variabile scelta. Nello specifico, "Z" è simultaneamente un determinante della variabile dipendente e correlato con il regressore.

Ciò può portare ad una sovra- o sotto-stima del coefficiente di regressione individuato.

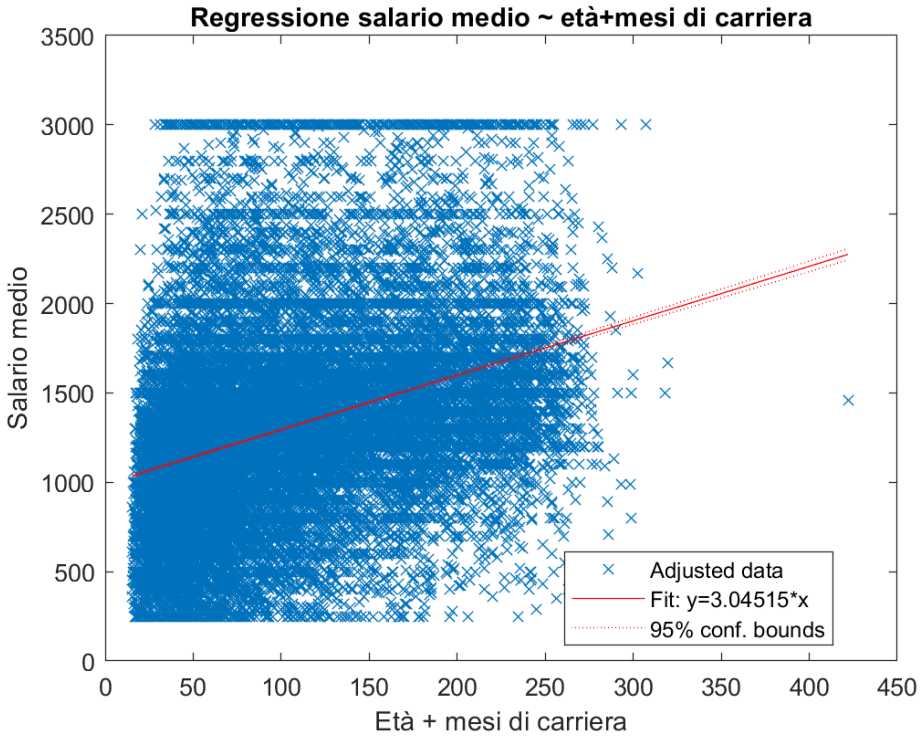
Tenendo conto del *bias*, sono state considerate diverse variabili (quantitative e qualitative) al fine di ottenere un modello di regressione più accurato.

Dall'analisi delle matrici di correlazione e covarianza e dall'osservazione delle distribuzioni complessive sono state selezionate, rispettivamente:

- il numero complessivo dei mesi di carriera lavorativa ("DURATT") come variabile quantitativa;
- il titolo di studio ("TISTUD") riqualificato come variabile binaria "con almeno titolo di laurea/senza titolo di laurea" , al fine di sottolineare eventuali differenze significative in termini di salario medio.

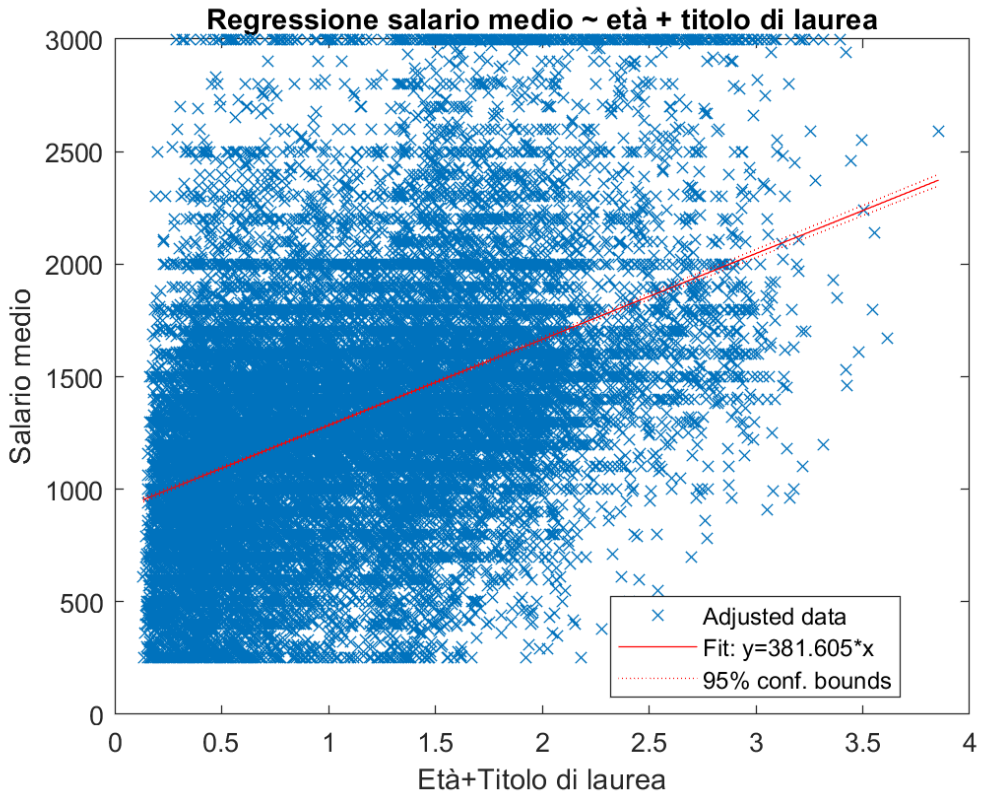
Entrambi i modelli elaborati hanno mostrato risultati migliori del primo modello bi-variato:

- Dall'integrazione della variabile "DURATT" la capacità di spiegare il fenomeno osservato è salita al 12.6%.



(Figura 5)

- Dall'integrazione della variabile "TISTUD" riqualificata la capacità di spiegare il fenomeno osservato è salita al 13.7%.



(Figura 6)

Ciò ha consentito un maggiore avvicinamento del modello di regressione alla realtà, rendendolo ulteriormente rappresentativo, anche se ancora lontano dall'affidabilità desiderata.

Regressione multivariata

In riferimento all'importanza di introdurre un maggior quantitativo di variabili all'interno del modello di regressione per ottenere una più efficace rappresentazione dell'andamento del salario medio mensile, dopo un'attenta analisi della matrice di correlazione per escludere multicollinearità, sono state identificate 7 ulteriori variabili, sia qualitative che quantitative, in grado di descrivere parzialmente la distribuzione del salario medio.

Questo nuovo modello vanta una maggiore potenza esplicativa , arrivando a spiegare più di 2/5 della variabilità dei dati analizzati (più precisamente il 44%).

In particolare, nel corso dell'analisi, è stato possibile verificare come si modifica il livello del salario medio rispetto all'introduzione delle suddette variabili, ognuna elaborata per evidenziare sottogruppi significativi:

- Dato l'incremento di un'unità dell'età (ETAM), si evince un aumento del livello del salario di circa 8 euro;
- avere un titolo di studio uguale o superiore alla laurea (TISTUD) comporta un aumento del livello del salario medio di circa 370 euro;
- essere residente in una regione del Nord (RIP5) determina un aumento del salario medio di circa 116 euro;
- essere un lavoratore di genere maschile (SG11) comporta un livello del salario medio più alto di circa 203 euro;
- un lavoratore con un contratto a tempo pieno (PIEPAR) riceve un salario medio più alto di circa 498 euro;
- per un lavoratore italiano (CITTAD) si evidenzia un incremento del livello del salario medio di circa 235 euro;
- avere un contratto di lavoro a tempo indeterminato (DETIND) comporta un salario medio più alto di circa 210 euro.

	Estimate	SE	tStat	pValue
(Intercept)	-79.399	13.031	-6.0931	1.1231e-09
x1	8.1564	0.23458	34.77	5.4792e-259
x2	369.86	6.2545	59.136	0
x3	115.67	4.9036	23.588	9.6778e-122
x4	203.26	5.1888	39.173	0
x5	498.49	6.4612	77.151	0
x6	235.36	7.612	30.919	3.4115e-206
x7	210.37	7.1573	29.392	7.676e-187

(Figura 7)

Relazioni non lineari

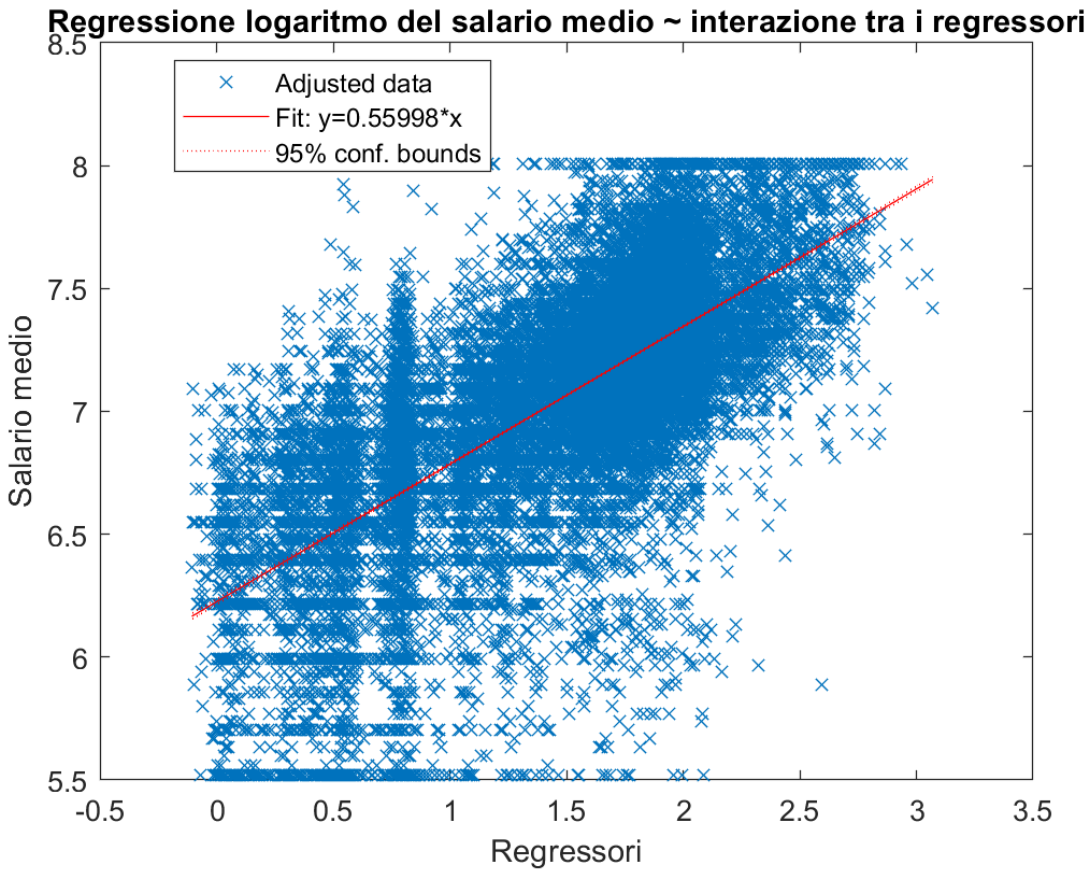
In ragione del fatto che la forza esplicativa del modello non risulti ancora ottimale, è stato fatto un tentativo di individuare una relazione non lineare in grado di spiegare in maniera più efficiente il comportamento del salario medio, al fine di raggiungere il minor livello di distorsione possibile. .

Osservando la nuvola di punti è stata individuata una relazione esponenziale, e si è cercato di catturare tale interazione utilizzando il logaritmo della variabile Y.

Modelli che utilizzano trasformazioni logaritmiche o polinomiali hanno mostrato risultati simili o leggermente migliori del modello lineare, ma quello che in particolar modo mostra una maggior adattabilità alla distribuzione di riferimento risulta essere una combinazione di una trasformazione logaritmica della variabile dipendente e una trasformazione interazionale dei regressori.

Como è possibile notare in figura, il modello mostra una migliore compatibilità con la distribuzione del salario medio, raggiungendo una efficacia illustrativa del 49.4% della variabilità totale della distribuzione,

.



(Figura 8)

Conclusioni

L'analisi condotta ha avuto l'obbiettivo si spiegare quali sono i parametri in funzione dei quali il livello di salario netto medio mensile subisce variazioni più o meno significative, al fine di ottenere un modello di regressione che fosse in grado di rappresentare la realtà in modo non distorto.

Durante le prime fasi di studio dei dati si sono analizzate le relazioni sussistenti tra il grado di formazione ed il livello di salario netto medio mensile, così come tra quest'ultima variabile e la regione di appartenenza. In riferimento allo studio di queste relazioni è stato possibile notare differenze significative tra chi ha un titolo di laurea specialistica a ciclo unico rispetto a chi ha altri titoli di laurea (triennale, diploma di Accademia, diploma universitario).

Allo stesso modo, è stato possibile notare una differenza significativa tra i salari medi dei rispondenti residenti in regioni del Nord rispetto a quelli dei rispondenti residenti in regioni del Sud.

Un primo modello di regressione bi-variata tra il salario netto medio mensile e l'età media ha evidenziato una leggera correlazione tra le due variabili, ma la forza esplicativa di tale modello risulta ancora troppo bassa.

Attraverso un'analisi della varianza dei residui è stato infatti notato come questa regressione bi-variata presenti il carattere di eteroschedasticità, e pertanto si è proseguito con l'elaborazione di nuovi modelli di regressione al fine di individuarne uno con una potenza esplicativa sufficientemente significativa.

Una prima ipotesi analitica è stata quella di aumentare la capacità del modello lineare inserendo nuove variabili qualitative e quantitative: nonostante tale processo abbia migliorato notevolmente la fruibilità del modello regressivo, la varianza dei dati e l'errore medio rimangono tali da impedire una considerazione adeguata dei dati osservati.

A tal proposito, la successiva ipotesi analitica è stata quella di individuare una relazione non lineare tra le variabili osservate: dopo numerosi tentativi il modello regressivo di riferimento è quello che evidenzia la relazione esponenziale (catturata trasformando in modo logaritmico il salario medio netto mensile) e una trasformazione interazionale tra i regressori individuati.

Il livello massimo di R^2 raggiunto è 0.494, che evidenzia come la capacità esplicativa e la performance del modello siano, seppur non ottimali, almeno soddisfacenti.

In definitiva, la variabilità dei dati forniti non consente un ulteriore miglioramento della modellizzazione del fenomeno studiato, ma possiamo concludere che le variabili utilizzate per descrivere il comportamento del livello del salario medio netto mensile abbiano una discreta capacità di spiegazione del fenomeno osservato.