Più istruzione non rende un Paese più ricco

Alessandro Pistola (1058248)
Corso di Laurea Magistrale in Informatica
Anno Accademico 2022-2023

Università di Bologna
Dipartimento di Informatica – Scienza e Ingegneria
Mura Anteo Zamboni 7
Bologna (BO), Italia

Indice contenuti

1 Introduzione

1.1 Descrizione del problema

La proposta di progetto si focalizza sull'analisi della relazione tra istruzione e crescita economica di un Paese (dove per crescita economica si intende la capacità di produzione, più in generale, di produrre ricchezza).
Comunemente si è portati a pensare che una forza lavoro ben istruita sia un fattore critico per lo sviluppo economico e ciò sarebbe dimostrabile dalla disparità tra il successo economico dei Paesi dell'Asia orientale, famosi per i loro alti livelli di istruzione, e la stagnazione dei Paesi dell'Africa subsahariana, che hanno livelli di istruzione tra i più bassi al mondo. Inoltre, l'ascesa dell'economia della conoscenza, in cui la conoscenza è diventata la principale fonte di ricchezza, ha reso l'istruzione, soprattutto quella universitaria, la vera chiave della prosperità.
Tuttavia, sono scarse le prove che più istruzione renda un Paese più ricco. Gran parte del "know-how" acquisito tramite l'istruzione in realtà non è determinante per l'incremento della produttività, anche se permette alla popolazione di avere una vita più soddisfacente e indipendente.
Inoltre, l'opinione che il sorgere dell'economia basata sulla conoscenza abbia determinato un aumento significativo del peso dell'istruzione può risultare fuorviante. Infatti, anche quando si tratta di educazione universitaria (che si suppone ancora più importante in un'economia della conoscenza) non c'è una rapporto automatico tra sapere e crescita economica.

1.2 Descrizione della soluzione proposta

La soluzione proposta è una soluzione nata da un processo di sviluppo comprendente operazioni di preprocessing dei dati, ricerca del modello ottimale e analisi statistica del modello.
Per lo sviluppo di tale progetto non si è fatto uso di uno specifico dataset, ma attraverso operazioni di manipolazione di dati sono state utilizzate più informazioni provenienti da diversi dataset pubblici.
La soluzione finale proposta è un notebook Jupyter per il primo modello sviluppato ed uno script R per il secondo modello proposto.
Le fasi sopracitate verranno introdotte, analizzate e discusse in dettaglio nei paragrafi seguenti.

2 Metodo proposto

2.1 Dataset

Come precedentemente accennato, il dataset si compone di diversi indicatori ottenuti da diversi dataset pubblici. Nei prossimi paragrafi verranno illustrate le operazioni eseguite per ottenere il dataset finale.

2.1.1 Acquisizione dati

Informazioni quantitative sull'istruzione
Dopo vari tentativi, scontrandosi con un problema costante di dati mancanti ed un problema di divisione dei Paesi (area geografica - Income - Suddivisione secondo metodo Pritchet [15]) si è scelto di partire dal dataset Barro e Lee [2] il quale assicura un buon punto di partenza con dati educativi quantitativi di 24 Paesi considerati avanzati e 122 Paesi in via di sviluppo (suddivisi ulteriormente in 6 grandi aree geografiche). A questo dataset si andranno poi ad aggiungere per i Paesi in questione altri indicatori qualitativi e indicatori economici preservando i raggruppamenti effettuati dagli autori del dataset.
Nello specifico il dataset contiene 13 campionamenti quinquennali per 146 Paesi. Ad ogni lustro sono associati i seguenti valori:

- **NoSchooling**: percentuale della popolazione over 15 che non ha frequentato nessun grado di istruzione.
- **PrimaryTotal**: percentuale della popolazione over 15 che ha terminato il ciclo di istruzione primario.
- **SecondaryTotal**: percentuale della popolazione over 15 che ha terminato il ciclo di istruzione secondario.
- **TertiaryTotal**: percentuale della popolazione over 15 che ha terminato il ciclo di istruzione terziario.
- **AvgYearsOfTotalSchooling**: Media degli anni di scolarizzazione della popolazione over 15.
- **Population(+15)**: popolazione over 15 espressa in migliaia.
- **Region**: regione di appartenenza.

Informazioni qualitative sull'istruzione
Al fine di inserire, seppur sia risultato di difficile modellizzazione, delle informazioni qualitative sull'istruzione dei Paesi, si sono utilizzati i risultati dei test TIMSS [1] e PISA [2]. A tal proposito, i dati sono stati estratti da Wikipedia per quanto riguarda gli score TIMSS e da OWID [12] per quanto riguarda gli score PISA. Sono stati poi generati i vari dataset (timss[anno edizione].xlsx e pisa.xlsx) e fuse le informazioni con il dataset BarroLee.

Spesa in educazione
I dati riguardanti la spesa in educazione (calcolata in percentuale del PIL/GDP) provengono da diverse fonti:

[1] Il Trends in International Mathematics and Science Study (TIMSS) dell'AIE è una serie di valutazioni internazionali delle conoscenze matematiche e scientifiche degli studenti di tutto il mondo.

[2] Programme for International Student Assessment è un'indagine internazionale promossa dall'OCSE nata con lo scopo di valutare con periodicità triennale il livello di istruzione degli adolescenti dei principali Paesi industrializzati. Comprende test per le materie lettura, matematica e scienze

- World Bank Data [20]: contiene la spesa in educazione come percentuale del PIL di ogni Paese, sono presenti dati dal 1970 al 2010.
- Institute of Education Sciences (Dati Stati Uniti) [6]
- Education For All In India (Dati India) [3]

Al fine di ottenere dati continui nell'intervallo 1970-2010, per ogni campione (Country) è stato selezionato per ogni lustro il valore dell'anno corretto (se presente) altrimenti il più recente nel quinquennio precedente.

Crescita economica
Come valori di crescita economica sono stati tenuti in considerazione i seguenti fattori:

- **GDP pro capite**: calcolato utilizzando il valore GDP reale di ogni anno calcolato con anno base 2017 (applicando i prezzi dell'anno base 2017 alla produzione dell'anno in considerazione) e la relativa numerosità della popolazione.
- **GDP per lavoratore**: calcolato come GDP pro capite (calcolato utilizzando il valore GDP reale di ogni anno calcolato con anno base 2017) (applicando i prezzi dell'anno base 2017 alla produzione dell'anno in considerazione) diviso la relativa numerosità della popolazione occupata.
- **GDP per ora lavorata**: calcolato come GDP per lavoratore diviso il numero medio di ore lavorate annualmente.

I dati di riferimento sono stati calcolati partendo dai dati forniti da Penn World Table [14].
Il dataset Pwt contiene dati per 183 nazioni con dati annuali dal 1950 al 2019.

A questo punto si è scelto di salvare su file il dataset contenente le informazioni fin qui presentate, essendo il più generale ed in grado di essere adattano a più analisi.
Le fasi di gestione dei valori mancanti, data transformation/normalization sono quindi specifiche per la tipologia di modello in esame.
Si è scelto di salvare il dataset finale nel file excel "final_dataset.xlsx"

2.2 Modelli proposti

L'obiettivo del progetto è identificare le influenze dell'istruzione sulla crescita economica, tenendo conto di diversi Paesi e dei cambiamenti temporalmente specifici all'interno di ciascun Paese. A tale scopo, la scelta del modello che si considera ottimale è quella dei modelli a effetti misti in quanto quest'ultimi consentono di sfruttare le informazioni provenienti da più Paesi o regioni simultaneamente e tiene conto delle variazioni specifiche dello specifico raggruppamento nel tempo.

Sono state considerate alternative come le semplici regressioni, ma queste non permettevano di utilizzare contemporaneamente le informazioni provenienti da più Paesi. Ciò limitava la possibilità di considerare l'eterogeneità tra Paesi e gli effetti specifici dei singoli Paesi sulla crescita economica. D'altra parte, i modelli VAR (Vector Autoregression) avrebbero potuto essere una scelta potenziale, ma richiedono una dimensione campionaria sufficientemente ampia per stimare correttamente i parametri e identificare le relazioni causali. Tuttavia, considerando il numero limitato di osservazioni disponibili per ciascun

Paese, il campione totale sarebbe risultato inadeguato per i modelli VAR.

La scelta di utilizzare un modello a effetti misti si basa su diverse fonti che evidenziano i vantaggi di tale approccio per affrontare problemi di natura simile [1]. Studi precedentemente condotti nel campo dell'econometria e dell'economia hanno dimostrato che i modelli a effetti misti consentono di catturare e spiegare efficacemente l'eterogeneità tra Paesi e regioni grazie agli effetti casuali dei fattori indipendenti [4] [9].

Nel contesto dell'utilizzo di modelli a effetti misti su dati panel, durante la gestione dei valori mancanti e la standardizzazione è importante considerare alcune caratteristiche fondamentali dei dati e delle trasformazioni applicate [5].

2.2.1 Gestione valori mancanti

I modelli a effetti misti richiedono dati completi e senza valori mancanti. Ciò significa che non possono essere presenti valori NaN nel dataset utilizzato per l'analisi. La presenza di valori mancanti potrebbe compromettere la validità dei risultati, poiché i modelli a effetti misti si basano sull'assunzione di dati completi per stimare correttamente i parametri.
Eliminare gli anni senza dati per ogni Paese potrebbe creare un dataset con serie storiche non allineate nel tempo. Questo potrebbe influenzare i risultati di una regressione e potrebbe portare a interpretazioni errate o bias nei risultati.
Una possibile soluzione potrebbe essere considerare l'interpolazione dei dati mancanti. Si possono utilizzare tecniche di interpolazione per stimare i valori mancanti in base ai valori vicini nel tempo o utilizzando modelli predittivi. Questo aiuterebbe a mantenere l'allineamento temporale delle serie storiche e potrebbe fornire una rappresentazione più coerente dei dati nel tempo.
Tuttavia, è importante notare che l'interpolazione introduce un certo grado di incertezza e potrebbe influenzare i risultati dell'analisi.
Al fine di scegliere le corrette modalità di selezione dei dati che poi influiranno sulla scelta del modello e dello studio statistico, si è proceduto analizzando i valori NaN per ogni anno, feature, Paese e regione.

Essendo nell'ambito della modellazione di serie storiche, si è scelto di eliminare (e quindi non utilizzare direttamente nel modello proposto) le features relative allo score PISA e TIMSS (percentuale NaN $> 80\%$).
Inoltre, i valori della feature *EducationalExpenditure* non presentano valori per gli anni antecedenti al 1970. Congiuntamente alla presenza di un'alta percentuale di valori mancanti in quegli anni per le altre feature, si è deciso di utilizzare solamente i campionamenti dal 1970 in poi. Invece, per i valori mancanti dopo il 1970, si è proceduto analizzando Paese per Paese. Qualora si riscontri che un determinato Paese abbia più del 50% dei valori presenti si procederà mediante interpolazione mentre i rimanenti saranno rimossi.
Analizzando la feature *GDPcapita* si è notato come solamente 3 Paesi presentassero valori mancanti, si è quindi optato alla rimozione di quest'ultimi (Afghanistan, Cuba e Ucraina).
Per la feature *GDPworker* mancando solamente uno/due valori per gli anni 1970 e 1975 di più Paesi si è proceduto attraverso un'interpolazione dei valori mancanti.
Infine, la feature *GDPHour* è stata rimossa in quanto presentava più del 50% di valori nulli ed i dati presenti riguardavano per lo più Paesi della regione "Advanced Economics".

2.2.2 Data transformation/normalization

Molti modelli econometrici fanno uso di forme logaritmiche per analizzare al meglio le relazioni tra le variabili [5].
Sebbene in primo luogo si fosse pensato che l'applicazione di **robust scaling** ai dati di input potesse risultare utile (soprattutto nelle successive fasi di analisi), considerando la possibile applicazione di un modello con una forma logaritmica si è scelto di non procedere alla standardizzazione.

3 Risultati sperimentali

3.1 Tecnologie utilizzate

Lo sviluppo dei modelli ad effetti misti per cercare di comprendere le interazioni e le relazioni tra le variabili indipendenti (istruzione) e dipendenti (crescita economica) è stato effettuato mediante lo sviluppo di due modelli, uno estensione dell'altro.
Per il primo modello proposto ad effetti fissi (+ random intercepts) si è utilizzato il linguaggio di programmazione Python e le librerie Scikit-learn [16], Pandas [13], NumPy [11], Statsmodels [17] e Matplotlib [8], mentre il modello a effetti fissi e casuali è stato sviluppato utilizzando il linguaggio di programmazione R e alcuni pacchetti specifici come lme4, lmerTest e sjPlot (tutti reperibili attraverso *The Comprehensive R Archive Network* [18]).

3.2 Mixed Linear Model Regression

In questo specifico contesto di ricerca, l'utilizzo di un modello a effetti misti si configura come una scelta appropriata e vantaggiosa. Attraverso l'impiego di un modello a effetti misti, si è in grado di utilizzare in modo efficace l'intero set di dati disponibili, beneficiando di una dimensione campionaria più ampia.
Inoltre, questa tecnica di modellazione consente di tener conto delle correlazioni intrinseche tra i dati provenienti da diversi Paesi e regioni di appartenenza, fornendo un'analisi più completa [19].
Il primo modello proposto è il modello ad effetti misti più semplice e comune in cui si ha una singola struttura di raggruppamento (cluster) per l'effetto casuale.
La funzione di specificazione del modello è la seguente, dove la tilde divide le variabili dipendenti da quelli indipendenti:

```
GDPcapita ~ NoSchooling + PrimaryTotal + SecondaryTotal +
TertiaryTotal + AvgYearsOfTotalSchooling * EducationalExpenditure - 1
```

L'operatore $*$ riguarda le interazioni moltiplicative e permette di includere anche le singole colonne che sono state moltiplicate insieme mentre specificando "-1" viene rimossa l'intercetta dal modello a effetti fissi, andando a forzare la funzione a passare per lo zero.
A questo punto si ha una variabile dipendente GDPcapita e si cerca di "spiegare"/catturare parte della variazione in quest'ultima attraverso il fitting delle variabili NoSchooling, PrimaryTotal, SecondaryTotal, TertiaryTotal, AvgYearsOfTotalSchooling, EducationalExpenditure, AvgYearsOfTotalSchoolingEducationalExpenditure come effetti fissi.
Tuttavia, la variabile di risposta avrà della variazione residua (cioè varianza non spiegata) associata ad uno degli effetti casuali (fattori di grouping) come l'appartenenza ad una specifica regione.
Usando gli effetti casuali, si sta modellando quella variazione non spiegata attraverso la varianza.
Nel concreto, specificando un effetto casuale non si sta più vincolando il modello ad avere la stessa intercetta per tutti i gruppi di Paesi presenti nell'intero set di dati.
In statsmodels è possibile specificare un fattore di grouping durante l'inizializzazione dello stesso:

```
sm.MixedLM(y_1, X_1, groups=df_1['Region'])
```

Va notato che ogni qual volta che si inserisce un effetto casuale, l'ipotesi iniziale va modifcata, infatti ora ci si sta chiedendo se esiste un'associazione tra le variabili indipendenti

e il GDPcapita dopo aver controllato la variazione nelle regioni.

Scelta la formula con cui esplicare il modello proposto, sono state valutate le 4 forme funzionali maggiormente adottate in econometria [5]: log-log, log-lineare, lineare-log, lineare-lineare.
In questo caso è stata adottata la forma funzionale log-lin per modellare la relazione tra le variabili indipendenti e la variabile dipendente, che rappresenta il GDP pro capite. Questa scelta è stata motivata da diversi vantaggi offerti dalla forma log-lin.
Prima di tutto, utilizzando una forma log-lin, si è in grado di catturare la natura esponenziale delle relazioni economiche. In particolare, una trasformazione logaritmica della variabile dipendente consente di interpretare i coefficienti come effetti percentuali sul GDP pro capite. Questo fornisce una misura più intuitiva e facilmente interpretabile dei rapporti di impatto delle variabili indipendenti sul livello di sviluppo economico.
Per giustificare la scelta della forma funzionale log-lin, è stato effettuato un confronto tra diversi modelli utilizzando i criteri di informazione di Akaike (AIC) e di Bayes (BIC). Tali criteri, basati sulla massima verosimiglianza, valutano la bontà di adattamento del modello e penalizzano i modelli più complessi. Nella nostra analisi, i valori AIC e BIC sono stati minimizzati scegliendo il modello log-lin come forma funzionale migliore. Nel grafico 1 sono riportati i test statistici F per la significatività generale del modello ed i criteri di informazione AIC e BIC per ognuna delle forme funzionali.

```
F stat log-log model: <F test: F=array([[52.41206623]]), p=1.0124814883430429e-61, df_denom=812, df_num=7>
Bic: 1751.41
Aic: 1709.04


F stat log-lin model: <F test: F=array([[235.88688139]]), p=7.97564752558263e-191, df_denom=812, df_num=7>
Bic: 1753.33
Aic: 1710.96


F stat lin-log model: <F test: F=array([[19.1212419]]), p=9.857769908441406e-24, df_denom=812, df_num=7>
Bic: 18269.16
Aic: 18226.79


F stat lin-lin model: <F test: F=array([[20.71074886]]), p=1.0102815202047229e-25, df_denom=812, df_num=7>
Bic: 18274.38
Aic: 18232.01
```

Figure 1: Test F e valori AIC e BIC per le 4 forme funzionali

Inoltre, è stata condotta un'analisi dei residui per valutare la bontà di adattamento del modello log-lin. É importante che i residui siano centrati intorno allo zero e presentino una distribuzione casuale. Questo indica che il modello è in grado di catturare in modo soddisfacente la variazione della variabile dipendente. Un grafico dei residui centrato intorno allo zero e con un aspetto casuale suggerisce che il modello log-lin è appropriato per descrivere la relazione tra le variabili indipendenti e il GDP pro capite. Nella figura 2 è riportato uno scatterplot degli errori residuali, nella figura 3 è riportato il grafico di stima della densità del kernel del residuo del modello e come ultimo il grafico QQ (Quantile-Quantile) 4 (utilizzato per rappresentare i quantili teorici di una distribuzione normale) e i quantili empirici dei dati osservati. Tutti i grafici sono utilizzati per valutare la normalità della distribuzione dei residui del modello. Graficamente si possono notare alcuni outlier ma la distribuzione sembra tendere a quella normale.

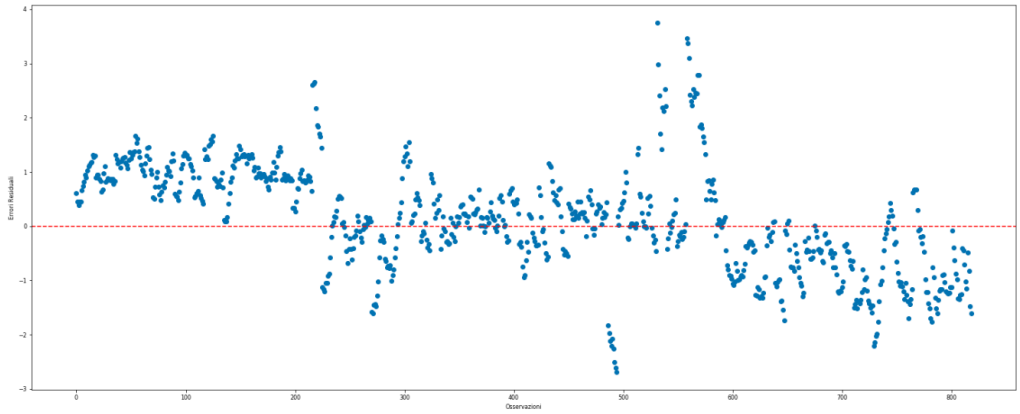

Figure 2: Grafico degli errori residuali

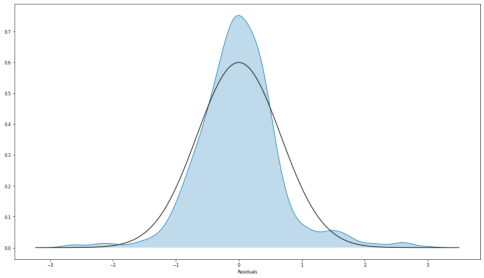

Figure 3: KDE Plot of Model Residuals

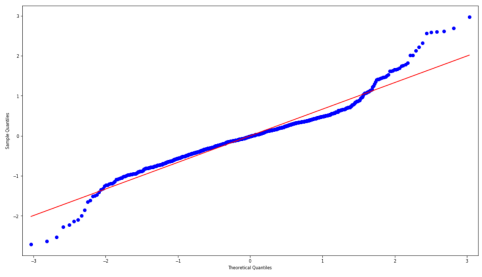

Figure 4: Q-Q Plot

3.3 Risultati

In questo paragrafo, nello specifico nella figura 5 sono riportati i risultati del fitting del modello che comprendono informazioni sul raggruppamento, che in questo caso coincide con l'effetto casuale e le stime di ogni coefficiente con relativi standard error, statistiche z e p-value ed intervalli di confidenza del 95%. Come constatabile dalla figura 5 tutti i

Model:	MixedLM	Dependent Variable:	np.log(GDPcapita)
No. Observations:	819	Method:	ML
No. Groups:	7	Scale:	0.4458
Min. group size:	9	Likelihood:	-846.4804
Max. group size:	216	Converged:	Yes
Mean group size:	117.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
NoSchooling	0.073	0.003	24.925	0.000	0.068	0.079
PrimaryTotal	0.066	0.004	17.681	0.000	0.058	0.073
SecondaryTotal	0.073	0.006	12.142	0.000	0.061	0.085
TertiaryTotal	0.062	0.009	7.203	0.000	0.045	0.079
AvgYearsOfTotalSchooling	0.319	0.050	6.383	0.000	0.221	0.417
EducationalExpenditure	0.130	0.039	3.366	0.001	0.054	0.206
AvgYearsOfTotalSchooling:EducationalExpenditure	-0.023	0.006	-4.068	0.000	-0.034	-0.012
Group Var	0.418	0.342				

Figure 5: Risultati mixed linear model regression con random intercepts

coefficienti sono significativamente diversi da zero per ogni livello di significatività (test Z), infatti nessun coefficiente presenta un p-value maggiore di 0.01 e nessun intervallo di confidenza contiene lo zero.

É possibile inoltre calcolare la quantità di variazione in GDPcapita spiegata dall'appartenenza ad una determinata regione. Per calcolare la percentuale di varianza spiegata dal raggruppamento (group-level variance), è necessario considerare la varianza totale dei dati e la varianza dei fattori casuali (group variance). La percentuale di varianza spiegata dal raggruppamento può essere calcolata dividendo la percentuale di varianza spiegata dal raggruppamento con la varianza totale dei dati.

$$VarExp(Region) = \frac{VarGroup}{VarGroup + scale} * 100 = 47\%$$

Quindi le differenze tra le regioni spiegano circa il 47% della varianza "rimanente" dopo aver escluso la varianza spiegata dagli effetti fissi.

3.3.1 Test di verifica d'ipotesi

Come test di verifica d'ipotesi si è utilizzato un test che potesse analizzare ipotesi multiple congiunte.
Il test utilizzato è il **Wald Test**. Il Wald Test è una prova statistica, usata per esaminare se un effetto esiste oppure no, esamina se una variabile indipendente ha un rapporto statisticamente significativo con la variabile dipendente.
Attraverso statsmodels è possibile esprimere l'ipotesi nulla sottoforma di stringa nel seguente modo:

```
hypothesis_0 = "EducationalExpenditure = 0, AvgYearsOfTotalSchooling = 0,
                AvgYearsOfTotalSchooling:EducationalExpenditure = 0"
```

attraverso la funzione wald_test è poi possibile calcolare il valore della statistica di Wald ed il relativo p-value:

```
test = log_lin_result.wald_test(hypothesis_0)
```

```
<Wald test (chi2): statistic=[[50.02284138]],
                   p-value=7.900192091798704e-11, df_denom=3>
```

Con un p-value pari a $7.9 * 10^{-11}$ l'ipotesi nulla è rifiutata per ogni livello α di significatività. É possibile quindi accettare l'ipotesi alternativa per cui non tutti i coefficienti sono statisticamente uguali a zero.

Sebbene visivamente, durante la scelta della forma funzionale si fosse verifica la forma normale dei residui del modello, questo non è mai stato dimostrato analiticamente.
Al fine di testare la normalità dei residui si è scelto il **test di Shapiro-Wilk** che è uno dei test più potenti per la verifica della normalità, soprattutto per piccoli campioni.
Tale test ha come ipotesi nulla (H_0) la normalità della distribuzione dei residui.
Di seguito vengono riportati i risultati del test:

```
Statistic 0.9491956233978271
p-value 3.547987817289973e-16
```

Al contrario di quanto verificato visivamente, con un p-value di $3.55 * 10^{-16}$ si rifiuta l'ipotesi nulla per ogni valore di α. Si dimostra quindi che l'assunzione sulla normalità degli errori non è rispettata.
Successivamente si è utilizzato un test di omoschedasticità per verificare l'assunzione di omoschedasticità dei residui (stessa varianza). Nello specifico si è utilizzato il **test di**

Breusch-Pagan. In questo caso, il test ha come ipotesi nulla H_0 l'omoschedasticità dei residui. Come sotto riportato, il test produce un p-value pari a $1.54 * 10^{-34}$ inferiore ad ogni livello di α. Si rifiuta quindi l'ipotesi nulla ed è possibile affermare che i residui presentano eteroschedasticità.
Ultima assunzione di particolare rilevanza in econometria ed in particolare nelle serie temporali è quella riguardante l'autocorrelazione dei residui. Si può avere un fenomeno di autocorrelazione temporale, a causa dell'inerzia o stabilità dei valori osservati, per cui ogni valore è influenzato da quello precedente e determina in parte rilevante quello successivo.
Il test utilizzato in questo caso è il **test di Durbin-Watson** che produce un valore di statistica pari a **0.34**. Valori agli estremi dell'intervallo [0, 4] indicano una correlazione, in questo caso essendo vicino al limite inferiore indica una correlazione positiva.

Riassumendo, seppur il test di Wald accerti la presenza di almeno un coefficiente statisticamente significativo, il modello viola le assunzioni di normalità e omoschedasticità dei residui e utilizza osservazioni autocorrelate positivamente.
Sebbene i coefficienti del modello siano statisticamente significativi, si ricorda che violazioni delle assunzioni possono influire sulla validità dei risultati del test di Wald e portare a conclusioni errate.

3.4 Approfondimento - effetti casuali

Precedentemente, introducendo i modelli a effetti misti, sono stati discussi solamente i modelli a effetti misti con intercetta casuale e si è affermato che sono i più semplici e comuni.
In questo paragrafo verranno introdotti gli effetti casuali *nested* e *crossed* per poi passare ai modelli ad effetti misti con random slopes.
Gli effetti casuali possono essere classificati come "nested" (annidati) o "crossed" (incrociati) a seconda di come sono strutturati all'interno del dataset. Gli effetti casuali annidati si riferiscono a situazioni in cui le unità di livello inferiore sono annidate all'interno delle unità di livello superiore. Ad esempio, i Paesi possono essere annidati ("nested") all'interno delle regioni.
Un effetto è (completamente) "crossed" quando tutti i soggetti hanno sperimentato tutti i livelli di quell'effetto. Nel contesto del seguente studio la variabile "Year" rappresenta proprio un effetto fully crossed infatti tutte le osservazioni di ogni Paese nell'anno 2010 potrebbero essere più simili tra loro perché hanno sperimentato le stesse stranezze economiche (recessione del 2008) piuttosto che perché vi siano stati investimenti maggiori in istruzione.
Inoltre, tali modelli ad effetti misti possono essere estesi includendo "random slopes" (pendenze casuali) che consentono alle relazioni tra le variabili indipendenti e dipendenti di variare tra le unità di livello superiore.
L'inclusione di effetti casuali annidati o incrociati e random slopes nei modelli ad effetti misti può migliorare la capacità di modellare la variazione dei dati e catturare le relazioni complesse tra le variabili. Questa flessibilità permette di tenere conto della struttura gerarchica o non gerarchica dei dati e di analizzare simultaneamente l'effetto delle variabili indipendenti a livello individuale e a livello di gruppo.
In questo modo, il modello terrà conto della non indipendenza nei dati: stessi Paesi sono stati campionati ripetutamente negli anni e più Paesi fanno parte di determinate regioni che possono avere diverse culture ed economie che influiscono su diversi fattori economici e dell'istruzione.

3.5 Random intercepts and random slopes

Nel capitolo precedente, il modello a effetti fissi con random intercepts è stato sviluppato utilizzando principalmente la libreria statsmodels di Python. Tuttavia, dopo innumerevoli tentativi con esito negativo in quanto non è possibile modellare effetti casuali troppo complessi, si è scelto di optare per l'utilizzo del linguaggio R il quale mette a disposizione pacchetti come lme4 e molti altri progettati con lo specifico scopo di sviluppare modelli a effetti misti.
Il secondo modello proposto è il modello ad effetti misti random intercepts e random slopes.
Per motivi di chiarezza si procederà incrementalmente partendo dalla formula di specificazione del modello a effetti fissi con random intercepts:

```
GDPcapita ~ NoSchooling + PrimaryTotal + SecondaryTotal +
TertiaryTotal + AvgYearsOfTotalSchooling * EducationalExpenditure - 1
```

In R utilizzando il pacchetto lme4 è possibile specificare più di un effetto, i fattori di grouping vengono specificato attraverso l'operatore "|", ad esempio con la seguente formula si ottiene lo stesso modello proposto nel capitolo precedente:

```
GDPcapita ~ NoSchooling + PrimaryTotal + SecondaryTotal + TertiaryTotal
+ AvgYearsOfTotalSchooling * EducationalExpenditure + (1|Region)
```

A questo punto si andrà ad aggiungere i due effetti casuali riportati come esempio nel paragrafo precedente, quello "nested" che riguarda l'appartenenza ad un determinato Paese di una specifica regione e quello "crossed" ovvero l'anno di campionamento. I due effetti casuali sono codificabili nella seguente forma:

```
(1 | Region/Country)
(1 | Year)
```

Aggiungendo i seguenti effetti casuali si permette all'intercetta di variare per ogni livello dell'effetto random, che tuttavia mantiene costante la pendenza tra quest'ultimi.
Quindi, in questo caso d'uso, l'utilizzo di questo modello significa che ci si aspetta che i Paesi in tutte le regioni mostrino la stessa relazione tra GDPcapita e fattori di istruzione/educativi negli stessi anni (pendenza fissa), sebbene si riconosca che alcuni Paesi potrebbero essere più ricchi o più poveri in partenza (intercetta casuale).
É facilmente ipotizzabile che non tutti i Paesi mostrino la stessa identica relazione, soprattutto se fanno parte di regioni differenti. Proprio per questo si procede alla codifica di un modello random intercepts e random slopes, in quanto è probabile che i Paesi delle aree più avanzate siano meno esposti ad esempio a variazioni in spesa educativa, mentre gli altri Paesi no.
Per consentire pendenze casuali è sufficiente aggiungere gli effetti fissi prima dell'operatore "|", in questo modo si permette alle variabili specificate di variare per i vari livelli dell'effetto casuale. I due nuovi effetti casuali sono quindi della forma:

```
(1 + NoSchooling + PrimaryTotal + SecondaryTotal + TertiaryTotal +
AvgYearsOfTotalSchooling * EducationalExpenditure|Region/Country)
(1 + NoSchooling + PrimaryTotal + SecondaryTotal + TertiaryTotal +
AvgYearsOfTotalSchooling * EducationalExpenditure|Year)
```

La funzione di specificazione completa del modello diventa la seguente:

```
log(GDPcapita) ~ NoSchooling + PrimaryTotal + SecondaryTotal +
TertiaryTotal + AvgYearsOfTotalSchooling * EducationalExpenditure +
(1 + NoSchooling + PrimaryTotal + SecondaryTotal + TertiaryTotal +
AvgYearsOfTotalSchooling * EducationalExpenditure|Region/Country) +
(1 + NoSchooling + PrimaryTotal + SecondaryTotal + TertiaryTotal +
AvgYearsOfTotalSchooling * EducationalExpenditure|Year)
```

Come per il modello a effetti fissi con random intercepts si riporta di seguito il risultato del fitting del modello a effetti fissi 6, il diagramma diagnostico dei residui 8 e la varianza catturata dagli effetti casuali 7.

```
Fixed effects:
                                                 Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                                     11.614627   2.928667 127.564597   3.966 0.000121 ***
NoSchooling                                     -0.029728   0.029629 104.859780  -1.003 0.318001
PrimaryTotal                                    -0.028092   0.029672 117.785492  -0.947 0.345697
SecondaryTotal                                  -0.020791   0.030476 109.924896  -0.682 0.496549
TertiaryTotal                                   -0.008258   0.032365  77.487476  -0.255 0.799278
AvgYearsOfTotalSchooling                        -0.029494   0.098073   5.533056  -0.301 0.774596
EducationalExpenditure                          -0.003348   0.028907   5.689300  -0.116 0.911775
AvgYearsOfTotalSchooling:EducationalExpenditure -0.001433   0.004533   4.304109  -0.316 0.766657
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) NSchln PrmryT ScndrT TrtryT AvYOTS EdctnE
NoSchooling -0.985
PrimaryTotl -0.981  0.991
SecondryTtl -0.950  0.966  0.987
TertiaryTtl -0.891  0.917  0.949  0.983
AvgYrsOfTtS -0.030 -0.032 -0.117 -0.258 -0.406
EdctnlExpnd -0.013 -0.090 -0.098 -0.123 -0.164  0.388
AvgYrOTS:EE  0.008  0.101  0.118  0.170  0.252 -0.597 -0.906
```

Figure 6: Risultati mixed linear model regression effetti fissi

```
Random effects:
 Groups         Name                                        Variance  Std.Dev. Corr
 Country:Region (Intercept)                                 2.159e-02 0.146942
                NoSchooling                                 1.491e-04 0.012211  0.35
                PrimaryTotal                                4.726e-04 0.021738 -0.07  0.49
                SecondaryTotal                              1.243e-03 0.035254 -0.55  0.32  0.86
                TertiaryTotal                               2.481e-03 0.049808 -0.55  0.42  0.81  0.98
                AvgYearsOfTotalSchooling                    1.435e-01 0.378762  0.48 -0.43 -0.83 -0.95 -0.97
                EducationalExpenditure                      1.141e-02 0.106814 -0.10 -0.33 -0.48 -0.35 -0.34  0.45
                AvgYearsOfTotalSchooling:EducationalExpenditure 2.261e-04 0.015037  0.08  0.43  0.61  0.48  0.48 -0.63 -0.93
 Year           (Intercept)                                 2.020e-02 0.142131
                NoSchooling                                 3.460e-06 0.001860 -1.00
                PrimaryTotal                                8.490e-06 0.002914 -0.84  0.83
                SecondaryTotal                              8.778e-06 0.002963 -0.92  0.92  0.98
                TertiaryTotal                               1.600e-06 0.001265 -0.52  0.53 -0.02  0.16
                AvgYearsOfTotalSchooling                    1.477e-04 0.012155 -0.04  0.03  0.57  0.41 -0.83
                EducationalExpenditure                      4.670e-04 0.021611  0.12 -0.14  0.43  0.26 -0.90  0.98
                AvgYearsOfTotalSchooling:EducationalExpenditure 1.634e-05 0.004043 -0.06  0.08 -0.48 -0.32  0.87 -0.99 -1.00
 Region         (Intercept)                                 1.758e-02 0.132572
                NoSchooling                                 1.264e-04 0.011243  0.57
                PrimaryTotal                                1.320e-04 0.011489  0.25  0.94
                SecondaryTotal                              3.203e-04 0.017898 -0.03  0.80  0.95
                TertiaryTotal                               8.134e-04 0.028521 -0.16  0.70  0.89  0.99
                AvgYearsOfTotalSchooling                    3.411e-02 0.184697  0.38 -0.51 -0.75 -0.91 -0.97
                EducationalExpenditure                      1.425e-03 0.037747 -0.50 -0.93 -0.89 -0.74 -0.66  0.51
                AvgYearsOfTotalSchooling:EducationalExpenditure 5.196e-05 0.007208  0.22  0.86  0.92  0.91  0.90 -0.81 -0.88
 Residual                                                   1.074e-02 0.103613
Number of obs: 819, groups:  Country:Region, 91; Year, 9; Region, 7
```

Figure 7: Risultati mixed linear model regression effetti casuali

Al contrario del modello a effetti fissi con random intercepts, il calcolo della *explained variance* è una questione molto complessa.
L'approccio utilizzato da Nakagawa e Schielzeth [10] è diventato l'approccio standard a questo tipo di problema (citato quasi 9000 volte).
Il pacchetto R MuMIn, incorpora gli algoritmi proposti da Nakagawa e Schielzeth [10] e quello di Johnson [7], un'estensione per permettere il calcolo anche sui modelli a random slopes.
Attraverso le funzioni esposte dal pacchetto è possibile calcolare il valore R^2 marginale

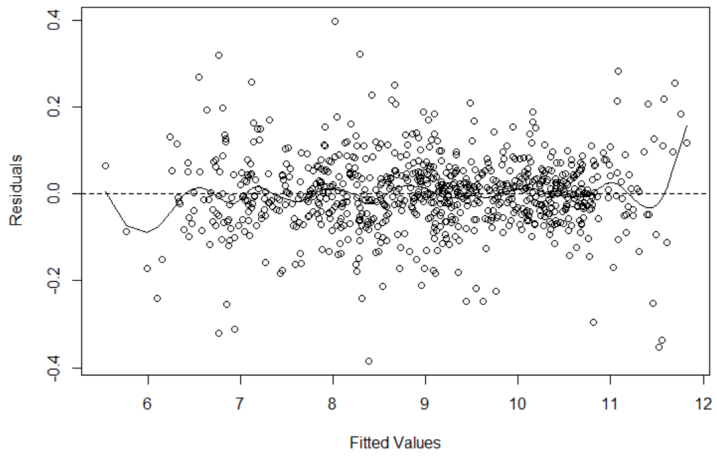

Figure 8: diagramma diagnostico dei residui

Tipologia modello	$R^2_{GLMM(m)}$	$R^2_{GLMM(c)}$	% Random effects
Fixed + random intercepts	27%	65%	58%
Fixed + random intercepts-slopes	3%	99%	97%

Table 1: Tabella di confronto per Pseudo-R-squared

e condizionale. L'R^2 marginale rappresenta la varianza spiegata daglie effetti fissi ed è calcolato come segue:

$$R^2_{GLMM(m)} = \frac{\sigma^2_f}{\sigma^2_f + \sigma^2_\alpha + \sigma^2_\epsilon}$$

L'R^2 condizionale rappresenta, invece, la varianza spiegata dall'intero modello ed è calcolata attraverso la l'equazione:

$$R^2_{GLMM(c)} = \frac{\sigma^2_f + \sigma^2_\alpha}{\sigma^2_f + \sigma^2_\alpha + \sigma^2_\epsilon}$$

La tabella 1 mostra gli R^2 marginali e condizionali per i 2 modelli proposti, evidenziando come l'utilizzo degli effetti casuali sia significativo. Un valore del 97% di varianza spiegata dagli effetti casuali sul totale della varianza spiegata dal modello indica che è stato effettuato un buon raggruppamento e che si stanno quindi rimuovendo tutti i fattori di variabilità comuni ai cluster.
Quindi, ad esempio, analizzando i Paesi appartenenti al cluster "Advanced Economies" il modello terrà in considerazione che il GDPcapita di quest'ultimi possa discostarsi di un certo valore da quello stimato dal modello a effetti fissi.
Nella figura 10 e 9 sono riportati gli effetti delle variabili indipendenti dei vari livelli degli effetti casuali "Region" e "Year" sui coefficienti del modello fisso.
Tuttavia, analizzando gli effetti fissi è possibile notare che l'unico coefficiente statisticamente significativo riguarda l'intercetta che rappresenta anche il fattore con più varianza catturata per gli effetti casuali "Year" e "Region".
In altre parole, sembrerebbe che il modello sia in grado di catturare molto bene le condizioni iniziali per ogni regione ed anno, ma la loro relazione con le altre variabili indipendenti non sembra essere significativa.

Figure 9: Random effects per Region

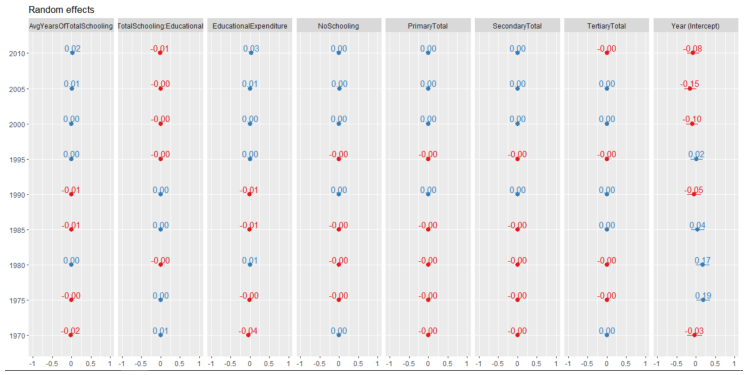

Figure 10: Random effects per Years

Il fatto che il modello a effetti fissi con random intercepts abbia un R^2 condizionale più basso ed un R^2 marginale più alto potrebbe significare che stia erroneamente utilizzando parte della variabilità causata dagli effetti casuali (poi modellati nel secondo modello proposto) nel modello a effetti fissi.
Questo potrebbe spiegare la significatività statistica dei coefficienti stimati dal modello ad effetti fissi con random intercepts del capitolo precedente.
É plausibile che il modello sviluppato in Python abbia sfruttato le variazioni e le correlazioni dovute ad osservazioni multiple dello stesso Paese (come già dimostrato dal **test di Durbin-Watson**) e le variazioni mondiali annuali nel GDPcapita (dovute alla variabile "Year"). Ora, utilizzando come effetto casuale il tempo, stiamo tenendo in considerazione le variazioni di crescita o decrescità delle variabili indipendenti e dipendente dovute allo sviluppo economico mondiale.
Ad esempio, se nell'anno 1980 l'Italia avesse registrato un aumento di cinque punti percentuali nel GDPcapita e nello stesso anno gli altri Paesi avessero registrato mediamente un aumento, il modello sarebbe in grado di tenere in considerazione tale fattore, non attribuendo, quindi, la variazione a fattori educativi degli anni precedenti.
Il modello random intercepts e random slopes indaga più a fondo la relazione di dipendenza del GDPcapita con gli indicatori educativi, in quanto tiene in considerazione più effetti casuali, attribuendo ovvero parte della varianza a quest'ultimi.
Per approfondire la relazione degli effetti casuali, nella figura 11 è riportato l'Average Marginal Effect di ogni variabile indipendente sulla variabile dipendente. Un AME negativo indica che, considerando l'effetto combinato di tutte le variabili indipendenti nel modello, un aumento nella variabile indipendente è associato a una diminuzione media

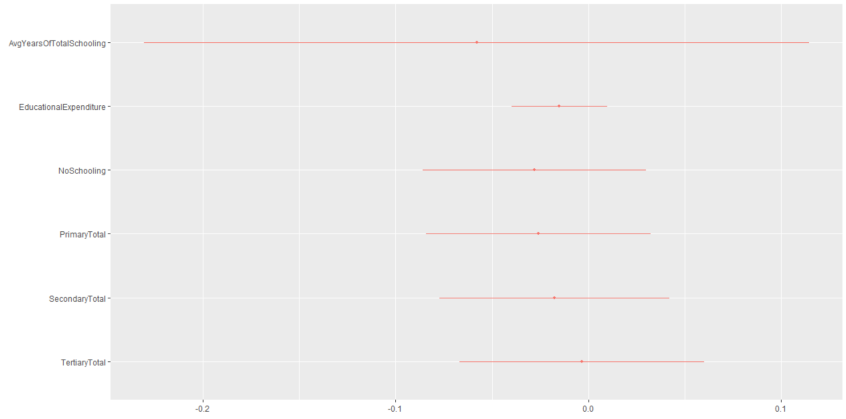

Figure 11: Random effects per Years

nella variabile dipendente. Questo implica che l'effetto medio della variabile indipendente sulla variabile dipendente, tenendo conto delle altre variabili, è negativo.
In primo luogo si potrebbe pensare ad un'associazione negativa tra i fattori educativi e quello economico, tuttavia è facile osservare che tutti i valori AME presentano un intervallo di confidenza comprendente lo zero.
Un intervallo di confidenza che comprende lo zero indica che non c'è evidenza statistica sufficiente per affermare che il parametro in questione sia diverso da zero. In altre parole, non possiamo affermare con sicurezza che il parametro abbia un effetto significativo o diverso da zero.
Come ultimi step si riportano i grafici delle predizioni marginali per le variabili *EducationalExpenditure* e *AvgYearsOfTotalSchooling* per regione e per anno (dove l'alone rappresenta l'intervallo di confidenza) figure 12, 13 e 14, 15; In conclusione, i coefficienti

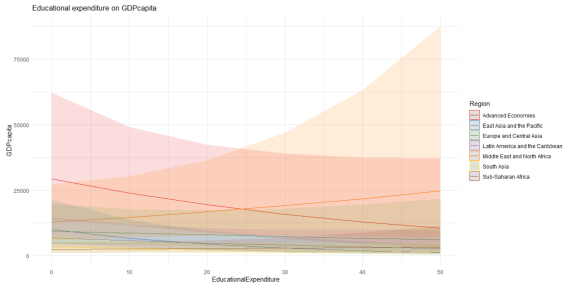

Figure 12: EduExpenditure/GDPcapita

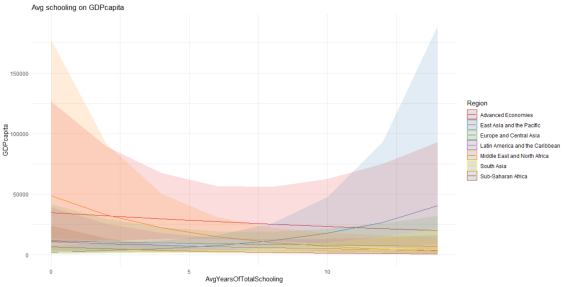

Figure 13: AvgYearsSchooling/GDPcapita

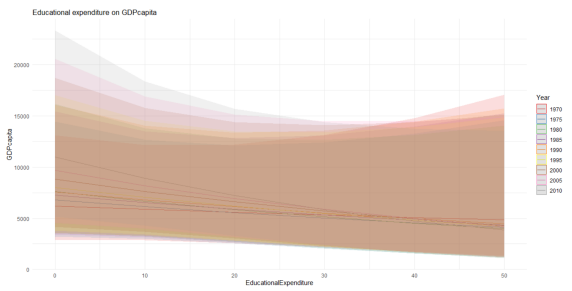

Figure 14: EduExpenditure/GDPcapita

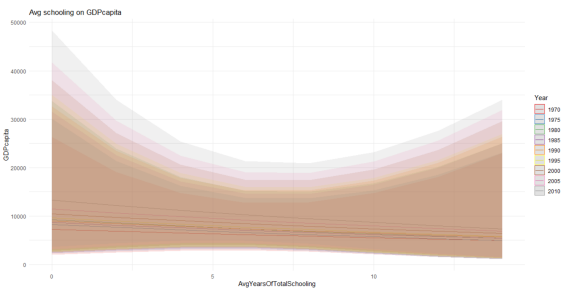

Figure 15: AvgYearsSchooling/GDPcapita

stimati dal modello a effetti fissi con intercetta e pendenza casuali rispondono al quesito posto in partenza riguardante la rilevazione e la quantificazione dell'eventuale associazione tra GDPcapita e fattori educativi dopo aver preso in considerazione le variazioni di GDP-capita annuali, per regione e intra-Paese.
Tuttavia, sebbene il modello a effetti fissi con random intercepts abbia dimostrato la presenza di un'influenza positiva della spesa in educazione e della media degli anni di scolarizzazione sul GDPcapita di un Paese, per le ragioni sopra citate, si ritiene tale risultato non affidabile.
Proprio per questo, si può concludere che da tale studio, utilizzando il dataset in questione, non è possibile rilevare una relazione statisticamente significativa tra GDPcapita (crescita economica) e fattori educativi.
I due grafici finali mostrano la relazione "media" degli effetti fissi e casuali tra le due variabili indipendenti *AvgYearsOfTotalSchooling* 17 e *EducationalExpenditure* 16 rispetto la variabile dipendente *GDPcapita*.

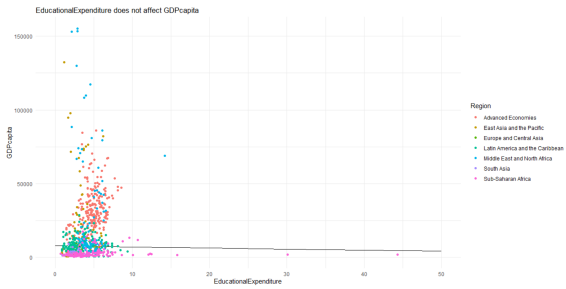

Figure 16: EduExpenditure/GDPcapita

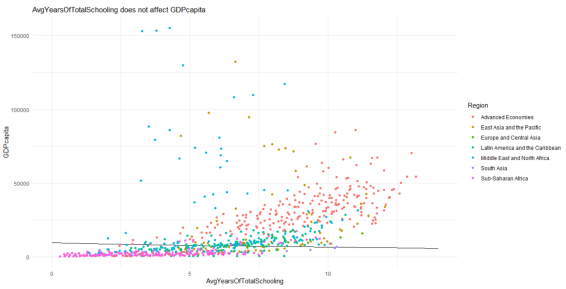

Figure 17: AvgYearsSchooling/GDPcapita

4 Discussione e conclusioni

4.1 Discussione dei risultati

Riassumendo, sono state applicate delle procedure di data manipolation e data preprocessing per creare un dataset che potesse risultare ottimale al lavoro di ricerca che si sarebbe dovuto svolgere nelle fasi succesive.
Dopo aver generato un dataset ritenuto informativamente completo, è stato proposto, in un primo momento, un modello a effetti fissi con intercetta casuale; una volta studiati ed evidenziati i limiti del modello, seppur questo sembrasse produrre risultati statisticamente significativi, sono stati introdotti i modelli a effetti misti con intercetta e pendenza casuale (random intercepts e random slopes). Tale modello, sviluppato utilizzando il linguaggio di programmazione R ha permesso di evidenziare ancor maggiormente le problematicità del primo modello.
Si è concluso lo studio visualizzando, attraverso grafici appositi, gli effetti peculiari prodotti dal modello a effetti misti, evidenziando, nel finale, la presupposta corretta formulazione del modello e la non evidenza statistica di una relazione tra la variabile dipendente e quelle indipendenti.
É quindi possibile concludere lo studio considerando l'obiettivo dello stesso raggiunto.

4.2 Limiti e lavori futuri

Sicuramente tra i limiti della soluzione proposta va menzionato il rischio di overfitting non citato finora.
Inoltre, sebbene i modelli a effetti misti consentono di gestire la presenza di variabilità sia tra le unità di livello superiore (gruppi) che all'interno delle unità di livello inferiore, molto spesso diventano iper-parametrizzati e complessi andando a rendere l'interpretazione dei risultati più complessa.
Le ultime due limitazioni che non possono non essere menzionate riguardano principalmente il dataset. Infatti, questo non presenta una elevata numerosità campionaria e i cluster gerarchici utilizzati per modellare gli effetti casuali non sono bilanciati.

Possibili continuazioni progettuali possono essere:

- Effettuare trasformazioni di scala (modificando la formulazione del modello) per indagare i motivi della non normalità dei residui.
- Utilizzare una procedura di model selection top-down come quella proposta da Zuur et al. [21].
- Tentare un approccio utilizzando metodi com modelli Generalized Linear Mixed Models.
- Ridurre la complessità del modello al fine di rendere lo stesso maggiormente interpretabile.
- Bilanciare il dataset di partenza e aumentare i campioni.
- Utilizzare tutte le informazioni tralasciate del dataset (GDPhour, GDPworker) compresi i dati qualitativi riguardanti l'educazione.

References

[1] Baltagi, B. H., & Baltagi, B. H. (2008). Econometric analysis of panel data (Vol. 4). Chichester: Wiley.

[2] Barro, Robert and Jong-Wha Lee, 2013, "A New Data Set of Educational Attainment in the World, 1950-2010." Journal of Development Economics, vol 104, pp.184-198. URL: http://www.barrolee.com/

[3] Education For All in India, URL: https://www.educationforallinindia.com/selected-educational-statistics-2000-2001.pdf

[4] Gałecki, A., Burzykowski, T., Gałecki, A., & Burzykowski, T. (2013). Linear mixed-effects model (pp. 245-273). Springer New York.

[5] Hill, R. C., Griffiths, W. E., & Lim, G. C. (2013). Principi di econometria. Zanichelli.

[6] Institute of Educational Sciences, URL: https://nces.ed.gov/pubs2016/2016014.pdf

[7] Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods in ecology and evolution, 5(9), 944-946.

[8] Matplotlib, python package, URL: https://matplotlib.org/

[9] McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. Psychological Methods, 24(1), 20.

[10] Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in ecology and evolution, 4(2), 133-142.

[11] NumPy, scientific computing library, URL: https://numpy.org/

[12] Dataset Our World In Data, Pisa score, URL: https://github.com/owid/owid-datasets/blob/4c73a1d2b4723c5a85d2370ef41ecc026129e643/datasets/OECD%20Education!%20PISA%20PISA%20(2015)/OECD%20Education!%20PISA%20Test%20Scores%20-%20PISA%20(2015).csv

[13] Pandas, data manipulation library, URL: https://pandas.pydata.org/

[14] Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), "The Next Generation of the Penn World Table" American Economic Review, 105(10), 3150-3182, available for download at www.ggdc.net/pwt

[15] Pritchett, L. (2001). Where has all the education gone?. The world bank economic review, 15(3), 367-391.

[16] Scikit-learn, URL: https://scikit-learn.org/stable/

[17] Statsmodels, URL: https://www.statsmodels.org/stable/index.html

[18] The Comprehensive R Archive Network, URL: https://cran.r-project.org/

[19] UCLA: Statistical Consulting Group. Introduction to linear mixed models. from https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

[20] Dataset World Bank Data, Educational Expenditure as GDP percetage, URL: https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS

[21] Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Mixed effects models and extensions in ecology with R (Vol. 574, p. 574). New York: springer.