Capitale umano

Introduzione

Quando si fa riferimento al termine "capitale", inteso in un senso generico, spesso tendiamo ad accostarlo al solo e semplice capitale fisico. Impianti, macchine, computer e ogni altro mezzo facente parte del processo produttivo rientrano certamente in questo grande insieme, eppure da soli non bastano a delinearne i confini.

Accanto al capitale considerato in termini meramente fisici ce n'è infatti un altro, dai contorni forse (anzi certamente) più fumosi ma non per questo meno importante: si tratta del capitale umano.

All'interno di questa nuova definizione rientrano tutte quelle conoscenze, abilità e competenze individuali sia innate che acquisite dai lavoratori attraverso l'istruzione e la formazione professionale. Ma non solo, *"il capitale umano comprende tutto ciò che influenza la capacità degli individui di produrre e creare reddito, oltre alla forza delle loro braccia: la salute fisica e mentale ne è una determinante fondamentale"**. Fermarsi solo all'istruzione potrebbe apparire quindi ad un primo sguardo riduttivo, rischiando di omettere importanti componenti del processo di sviluppo della persona e del lavoratore quali possono essere ad esempio i fattori sanitari, fondamentali per interpretare in una maniera il più esaustiva possibile questa diversa accezione di capitale.

Oggetto di questo lavoro sarà tuttavia principalmente la relazione tra la prima definizione qui data di investimento in capitale umano e il tenore di vita nei diversi Paesi del mondo (misurato attraverso il PIL pro capite a parità di potere d'acquisto). Prendendo a riferimento i dati relativi a 220 Paesi (anno 2015) pubblicati dalla Banca Mondiale in tema di spesa per istruzione, formazione e sviluppo si andrà quindi ad analizzare se ed in quale misura questo "investimento nell'individuo" possa portare ad un benessere dal punto di vista economico allo Stato che se ne fa promotore, il tutto con l'obiettivo costante di seguire un percorso il più possibile coerente e lineare.

** Intervento del Governatore della Banca d'Italia Ignazio Visco in occasione dei 15 anni di attività della Facoltà di Economia dell'Università Cattolica, sede di Roma*

Statistica descrittiva

Uno squardo al Pil pro capite

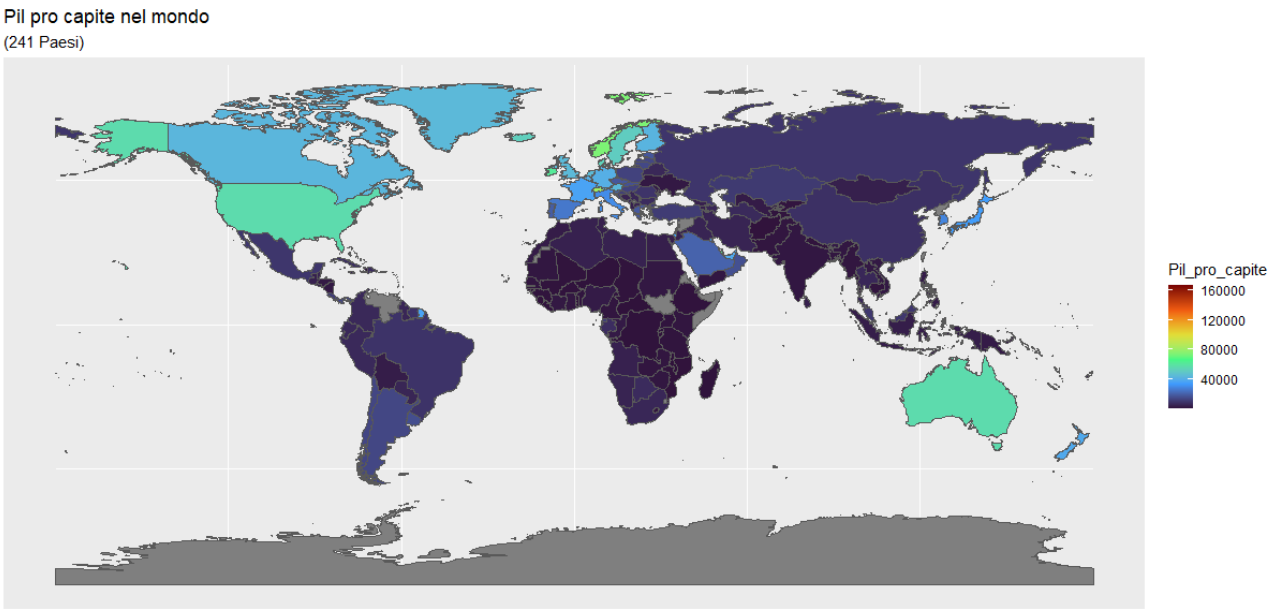
Prima di iniziare l'analisi dell'influenza esercitata dall'investimento in capitale umano sul tenore di vita nei diversi Paesi è giusto focalizzarsi sulla variabile risposta di questa ricerca: il Pil pro capite, inizialmente non considerandolo a parità di potere d'acquisto.

Come si distribuisce questa variabile a livello mondiale?

Si può osservare che la distribuzione del Pil non appare per nulla omogenea: infatti, ad una situazione caratterizzata da livelli di reddito pro capite molto bassi (soprattutto in Africa, Sud America e Asia), se ne contrappone un'altra speculare in cui il tenore di vita risulta essere decisamente più elevato (Europa, Nord America e Oceania).

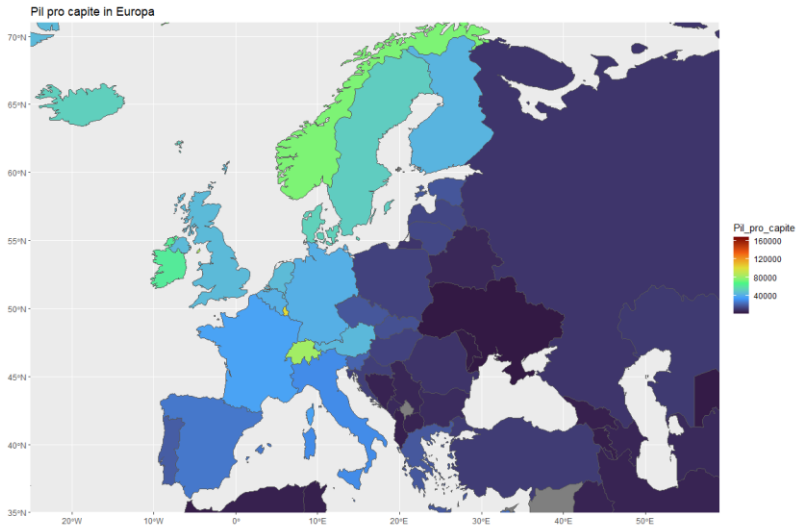
Di particolare rilevanza sono i cosiddetti Microstati, ovvero quelle Nazioni aventi una popolazione ridotta e un territorio limitato, i quali occupano tutte le prime posizioni della classifica della distribuzione del Pil pro capite a livello mondiale. Difatti, il tenore di vita più elevato si ha nel Liechtenstein, seguito dal Principato di Monaco, dalle Bermuda, Lussemburgo e Isola di Man (tutte con un reddito pro capite maggiore di 85.000$).

All'opposto, i Paesi dell'Africa centro-meridionale (Burundi, Rep. Centrale Africana, Malawi, Madagascar e Niger) risaltano all'occhio per i bassi standard quali-quantitativi del tenore di vita, con un livello di reddito pro capite addirittura al di sotto di 490$.

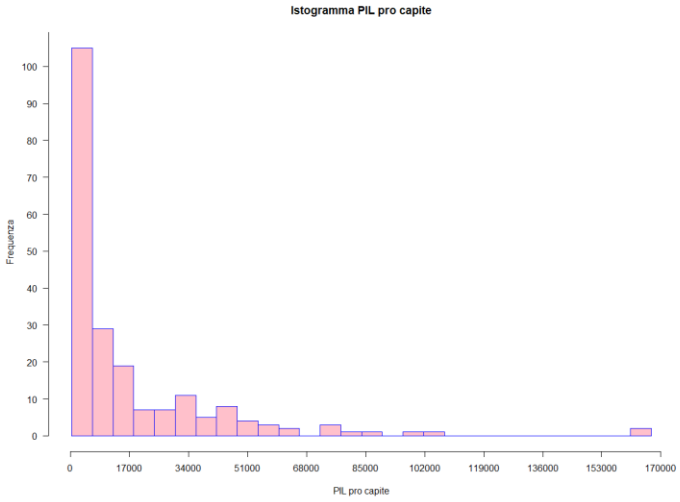


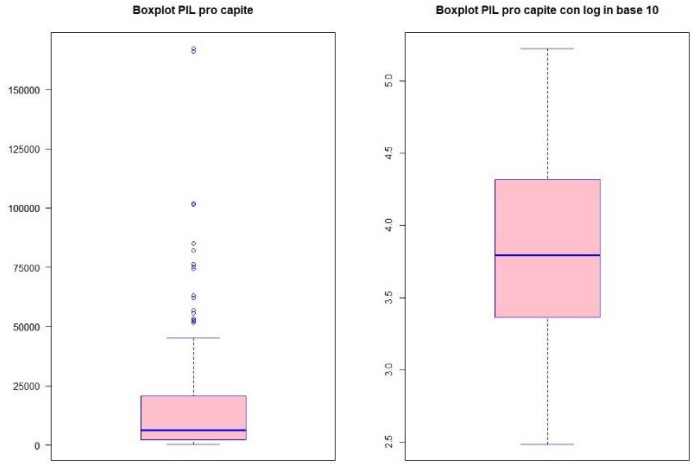
A livello europeo il tenore di vita risulta invece essere decisamente più omogeneo che a livello mondiale, sebbene sia comunque presente una differenza tra i Paesi del Nord e del Sud Europa. In particolare si distinguono per gli elevati livelli di Pil pro capite i Paesi scandinavi (insieme alla Svizzera, Irlanda e ai Microstati), seguiti dai Paesi dell'Europa continentale e Meridionale, mentre i Paesi dell'area Balcanica rimangono ancorati a bassi livelli di reddito.

L'Italia, con un Pil pro capite di circa 30.200$, si rivela essere uno degli Stati con gli standard più elevati nel Sud Europa, sebbene permangano delle leggere differenze con le Nazioni dell'area Continentale quali Francia (36.638$), Germania (41.086$), Austria (44.178$) e Regno Unito (44.974$).



I grafici riportati di seguito mostrano come la distribuzione del reddito per abitante non abbia un profilo per nulla simmetrico e, anzi, presenti un'evidente asimmetria positiva. Il 50% dei Paesi si colloca infatti tra un livello di Pil pro capite compreso tra i 2.132,2$ e i 17.879,5$, rendendo così le frequenze delle classi di reddito più basse sia quantitativamente che visivamente più rilevanti delle classi più alte.



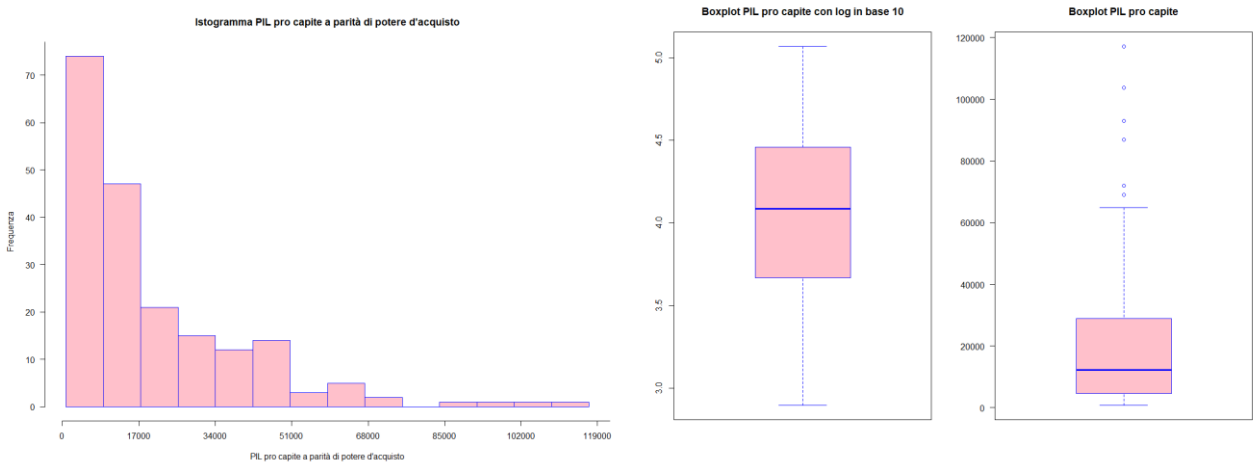


Il boxplot della distribuzione a livello mondiale del Pil pro capite: si evidenzia come la maggior parte dei Paesi si concentri in fasce basse di reddito per abitante, mentre spostandosi verso l'alto sia presente un elevato numero di valori anomali dato soprattutto dai Microstati. A fianco un boxplot realizzato utilizzando il logaritmo in base 10 dei livelli di Pil pro capite.

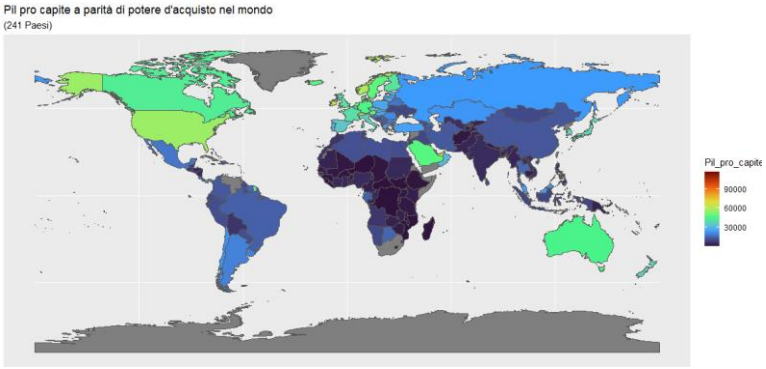
Focus: il Pil pro capite a parità di potere d'acquisto

Presentato il modo di distribuirsi del Pil pro capite tra i vari Paesi del mondo, si anticipa che nel corso della trattazione verrà utilizzata una versione particolare di questa variabile: il Pil pro capite a parità di potere d'acquisto. Questa misura tiene infatti conto delle differenze nel livello generale dei prezzi tra i diversi Stati.

Anche per questa variabile, per completezza, vengono presentati i grafici già analizzati nello studio del Pil pro capite.

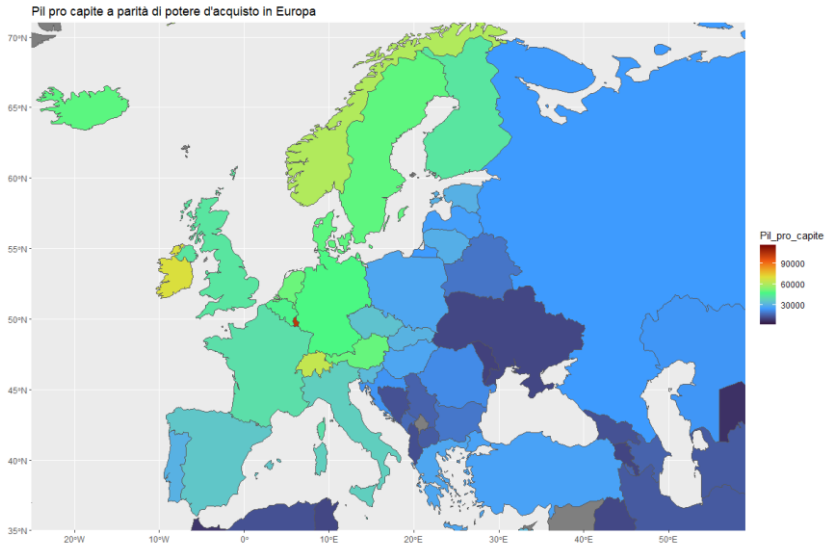


Istogramma e boxplot (anche in logatimi in base 10) del Pil pro capite PPA



Distribuzione a livello mondiale del Pil pro capite a parità di potere d'acquisto

Distribuzione a livello europeo del Pil pro capite a parità di potere d'acquisto



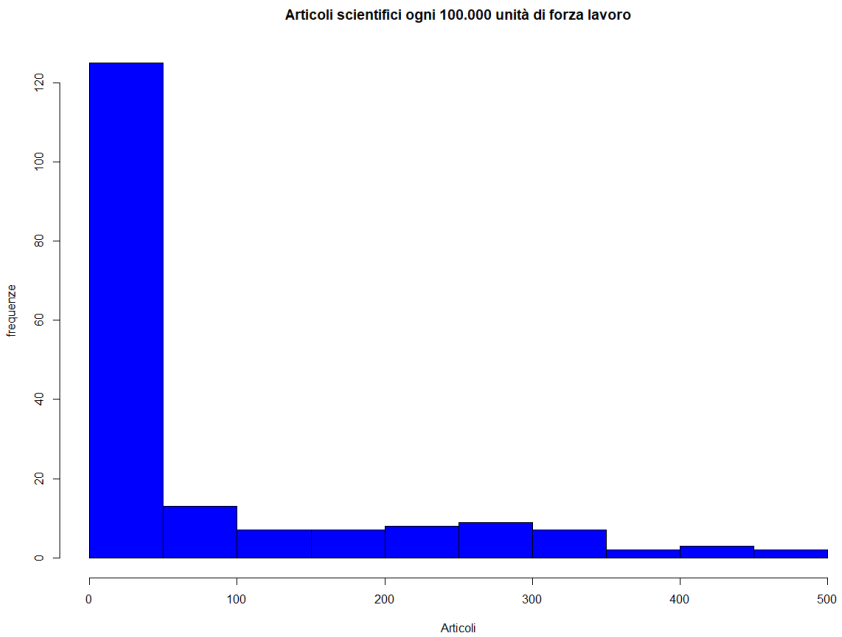
Due misure dell'investimento in capitale umano

Si intendono ora presentare due variabili rappresentative del livello di investimento in capitale umano nei diversi Paesi del mondo: *gli articoli scientifici ogni 100.000 unità di forza lavoro* e il *rapporto alunni-docenti nella scuola primaria*. Poiché queste misure ricorreranno nel corso della trattazione, verranno di seguito sinteticamente analizzate le loro principali caratteristiche e la loro distribuzione a livello mondiale.

Gli articoli di riviste scientifiche e tecniche

La pubblicazione scientifica è una misura importante nel tentativo di spiegare il differente contributo di ciascun Paese allo sviluppo del settore della ricerca e all'innovazione tecnologica, ma non solo, essa riflette anche l'impegno dei diversi Stati volto all'incremento del pool di conoscenze scientifiche e di competenze tecniche all'interno dei propri confini. Per "articoli di riviste scientifiche e tecniche" si intende così il numero di articoli scientifici e ingegneristici pubblicati nei seguenti campi: fisica, biologia, chimica, matematica, medicina clinica, ricerca biomedica, ingegneria e tecnologia e scienze della terra e dello spazio.

Poiché tale misura privilegiava però in una qualche maniera gli Stati più popolosi a danno di quelli con una popolazione ridotta, si è deciso di rapportare tale variabile alla forza lavoro presente nei diversi Paesi e successivamente di moltiplicarla per 100.000. Mentre infatti la grandezza originaria assumeva i suoi valori più elevanti in corrispondenza di Usa (circa 429.988 articoli), Cina (circa 407.974 articoli), Germania, India e Giappone, questa nuova misura tiene conto delle differenze riguardanti l'entità della popolazione nel mondo del lavoro nei diversi contesti di analisi, presentando così valori molto elevati per Paesi come la Danimarca (circa 492 articoli ogni 100.000 unità di forza lavoro), la Svizzera, l'Australia, la Finlandia e la Svezia. Inoltre, lì dove la variabile originaria presentava i suoi valori più bassi in presenza di Stati insulari con una popolazione molto ridotta (ma non per questo con un ridotto impegno nella ricerca scientifica), gli articoli scientifici ogni 100.000 unità di forza lavoro assumono i loro valori minori in corrispondenza del Sudan del Sud, della Somalia, del Chad, della Nord Corea e dell'Afghanistan, che contribuiscono proporzionalmente di meno allo sviluppo e all'innovazione tecnologica.



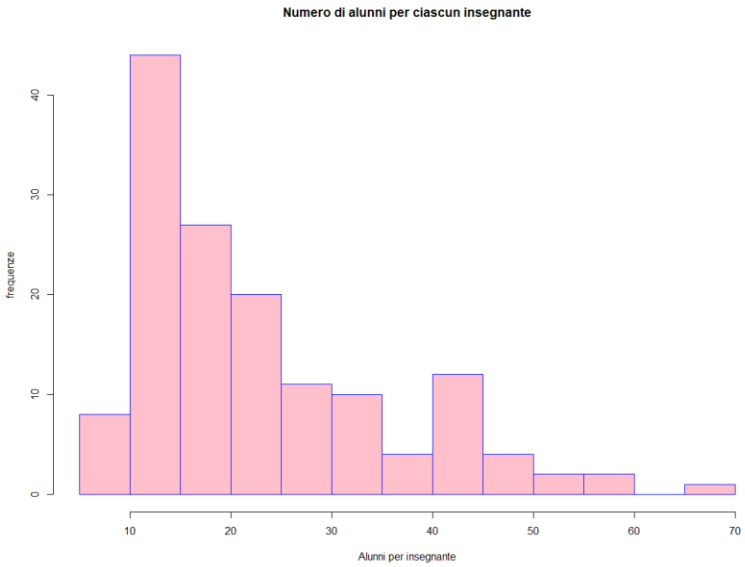
Si evidenzia anche qui, così come in precedenza nell'analisi della distribuzione del Pil pro capite, una evidente asimmetria positiva.

Difatti su un gruppo di 220 Paesi più della metà si trova nella classe 0-50, evidenziando così una difformità significativa, sebben intuibile, nell'apporto alla ricerca e allo sviluppo tecnologico dei differenti Paesi a livello mondiale.

Istogramma di frequenza del numero di articoli scientifici ogni 100.000 unità di forza lavoro

Il rapporto alunni-docenti nella scuola primaria

L'ultima grandezza che si intende presentare prima di iniziare la trattazione riguardo l'inferenza statistica è il rapporto alunni-docenti nella scuola primaria, ovvero il numero medio di alunni per singolo insegnante nei vari Paesi del mondo. Avendo presentato in precedenza una misura del contributo e dell'investimento dei diversi Stati nel settore della ricerca si tenta ora, mediante tale rapporto, di dare un'immagine della qualità dei differenti sistemi scolastici. Il focus sarà così indirizzato verso una particolare tipologia di investimento in capitale umano: quello verso i primi gradi di istruzione.

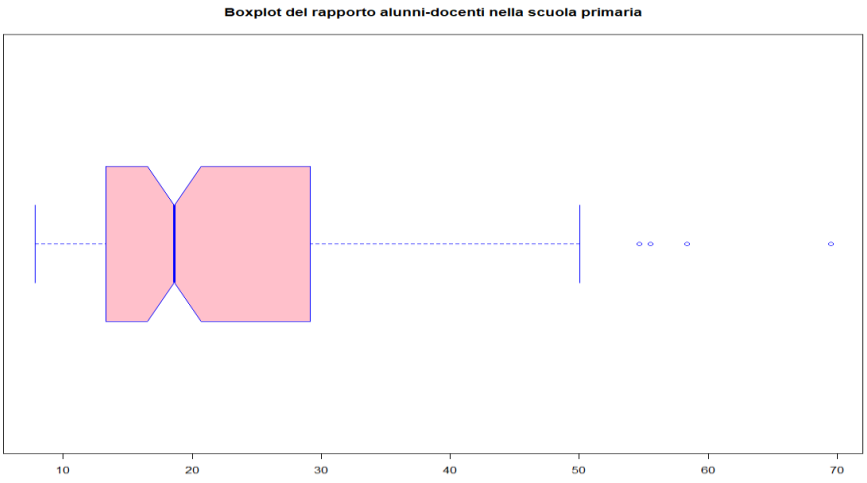


Anche tale distribuzione presenta, come le due precedenti, un'asimmetria positiva, avendo la coda tendente verso destra. L'interpretazione che se ne dà è tuttavia differente: mentre in precedenza bassi livelli di Pil pro capite o di pubblicazioni scientifiche segnalavano un certo livello di inefficienza, qui un ridotto rapporto alunni-insegnanti è invece sintomo della qualità del sistema scolastico a livello primario dei singoli Paesi.

Istogramma di frequenza del numero di alunni per insegnante nella scuola primaria

Le Nazioni con un più elevato rapporto alunni-docenti risultano così essere ancora una volta quelle del continente africano, tra cui il Malawi (con una media di 69,5 alunni per insegnante), il Rwanda (con una media di circa 58,3), il Chad, il Mozambico o l'Angola. All'opposto il rapporto tende ad assumere i suoi valori minimi in presenza soprattutto di Microstati come il Liechtenstein e il Lussemburgo (entrambi con una media minore di 8,5 alunni per docente) o dei Paesi del Nord Europa tra cui ad esempio la Norvegia (con circa 8,9 alunni per docente).

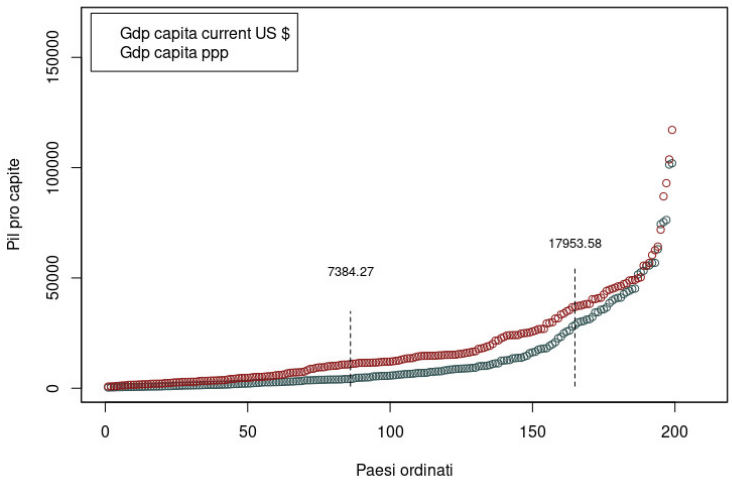
Il boxplot riportato qui di fianco evidenzia anch'esso l'asimmetria positiva già evidenziabile dall'istogramma. Il 50% dei dati è compreso nell'intervallo 13,3-29,1.



Statistica inferenziale

Pil pro capite e Pil pro capite a parità di potere d'acquisto a confronto

Come anticipato nella parte riguardante la statistica descrittiva, nel corso della trattazione si utilizzerà una misura particolare del Pil pro capite, ossia quella che tiene conto del livello generale dei prezzi nei vari Paesi del mondo. Si intende così introdurre questa nuova fase del lavoro mediante un breve confronto tra il Pil pro capite e il Pil pro capite a parità di potere d'acquisto.



Il grafico a sinistra mette così in evidenza l'andamento delle due diverse variabili a partire dai Paesi più poveri sino a quelli tendenzialmente più ricchi, identificando con il colore verde il *Pil pro capite* e con il colore rosso il *Pil pro capite PPP.* L'andamento delle due misure risulta essere il medesimo, sebbene il Pil pro capite a parità di potere d'acquisto assuma valori leggermente più elevati soprattutto in corrispondenza di Stati con un livello medio di reddito per abitante.

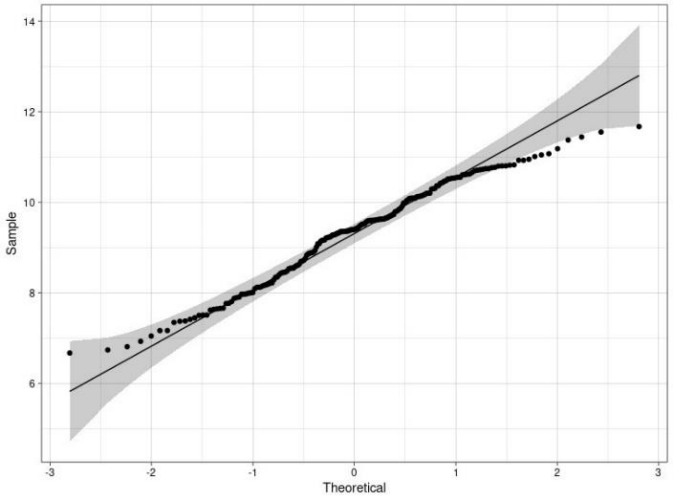
Un'ultima assunzione rilevabile dal grafico: si è deciso dividere le varie Nazioni oggetto dello studio in tre categorie, ossia in Paesi poveri (fino a 7.384,27$ di Pil pro capite PPP), Paesi intermedi (7.384,27$-17.953,58$) e Paesi ricchi (più di 17.953,58$): nel seguito verranno così utilizzati i Paesi poveri nell'ambito della stima bayesiana e i Paesi medio-ricchi nel modello di regressione.

La distribuzione del Pil pro capite PPP: test di Shapiro-Wilk e test di Kolmogorov-Smirnov

Volendo analizzare quale potesse essere la forma della variabile casuale generatrice dei dati campionari in nostro possesso riguardanti il Pil pro capite PPP si è inizialmente voluta testare l'ipotesi di provenienza da una popolazione normale tramite il test di Shapiro-Wilk ed una contestuale verifica grafica tramite il qqplot. Pur avendo espresso la variabile di nostro interesse in logaritmi (tentando così di ridurre le differenze tra Paesi ricchi e poveri), il test non ha dato esito positivo.

```
        Shapiro-Wilk normality test

data:  log(gdp_capita)
W = 0.97946, p-value = 0.005142
```

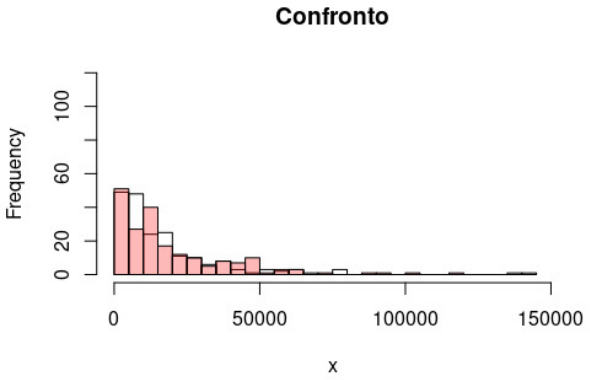
Output di R del test di Shapiro-Wilk: il p-value risulta minore di 0,05 per cui si rifiuta l'ipotesi nulla di distribuzione normale della popolazione. Di fianco il qq-plot, nel quale si può osservare come i punti divergano dalla diagonale, specialmente nelle due code.

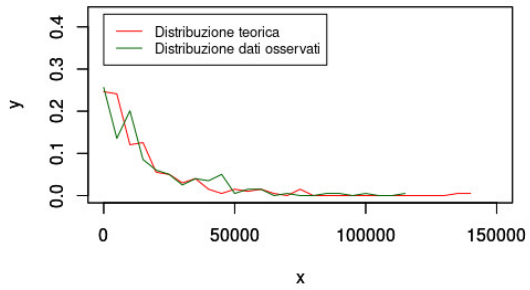


Rifiutata l'ipotesi di provenienza da una popolazione normale e ricordata l'asimmetria positiva nella distribuzione del reddito pro capite PPP osservata precedentemente, l'analisi si è così focalizzata sulla verifica della provenienza del campione da una variabile casuale Gamma. Per poter testare questa nuova ipotesi si è però reso necessario calcolare dapprima i parametri della variabile Gamma che si è supposto poter essere la generatrice dei nostri dati. Sono così stati calcolati i parametri shape e scale, i quali risultano essere rispettivamente *il rapporto tra media al quadrato e varianza del Pil pro capite PPP* (pari a 0,9307997) *e il rapporto tra varianza e media* della medesima variabile (pari a 20.355,12).

Per la verifica della provenienza del campione da una variabile casuale Gamma è stato inizialmente utilizzato il test di adattamento di Kolmogorov-Smirnov, il quale tuttavia non ha dato un esito positivo, sebbene le due distribuzioni sembrassero in prima analisi molto simili. Poiché il problema pareva essere dato più che altro dal basso numero di osservazioni in nostro possesso, dopo aver generato un campione casuale dalla distribuzione Gamma con i parametri stimati in precedenza, si è proceduto così ad un nuovo test di Kolmogorov-Smirnov per testare l'identica distribuzione delle popolazioni generatrici dei due campioni: tale test ha dato esito positivo. Per sicurezza il test è stato ripetuto confrontando i dati in nostro possesso con campioni generati casualmente da altre variabili casuali, ma in tutti questi casi i bassi livelli del p-value hanno portato a rifiutare l'ipotesi di identica distribuzione delle popolazioni.

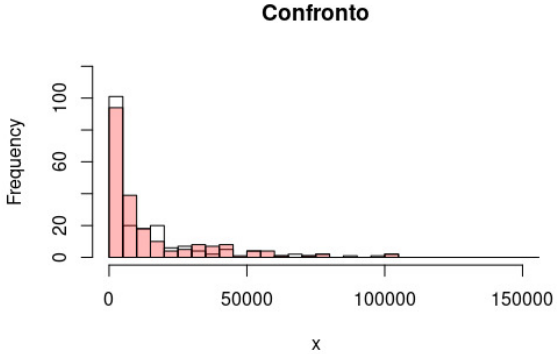
```
        Two-sample Kolmogorov-Smirnov test

data:  gdp_capita and x
D = 0.063207, p-value = 0.4085
alternative hypothesis: two-sided
```

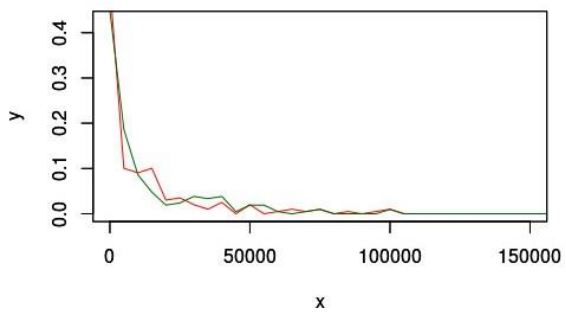




In alto a sinistra l'output di R relativo al test di Kolmogorov-Smirnov sull'identica distribuzione delle due popolazioni. In alto a destra un confronto tra l'istogramma di frequenze dei dati osservati del Pil pro capite PPP (in rosso) e del campione generato casualmente da una Gamma (in bianco). A sinistra un ulteriore confronto tra le due distribuzioni.

Si osserva inoltre nei due grafici sotto riportati che anche nel confronto tra la variabile "Pil pro capite" (non PPP) ed un campione casuale generato da una Gamma (con parametri shape e scale stimati utilizzando la medesima formula) si presenta una similitudine tra le due distribuzioni.





Stima puntuale

Dopo aver analizzato la distribuzione della popolazione di riferimento dei dati in nostro possesso, il prossimo tema ad essere presentato sarà quello della stima puntuale. Verranno così introdotti separatamente gli stimatori BLUE di media e varianza e gli stimatori MLE dei parametri shape e scale per i Paesi medio-ricchi e per i Paesi poveri, presentando poi per questi ultimi un approccio di stima puntuale di tipo bayesiano.

Gli stimatori BLUE di media e varianza per i Paesi medio-ricchi

Si presentano inizialmente gli stimatori BLUE di media e varianza di quella fascia di Paesi in precedenza definiti medio-ricchi (Pil pro capite PPP maggiore di 7.384,27$). Poiché tali stimatori presumono unicamente la conoscenza del campione, non è stata fatta alcuna congettura riguardo la distribuzione della popolazione da cui provengono i nostri dati.

Di fianco viene riportato l'output di R contenente gli stimatori BLUE della media e della varianza nella popolazione; essendo inoltre tali stimatori non distorti, ne vengono anche presentate le varianze (terza riga e quarta riga) che in questo caso coincidono con gli errori quadratici medi.

```
> (BLUE_mean <- BLUE_est$muhat)
          [,1]
[1,] 31278.01
> (BLUE_sigma <- BLUE_est$sigmahat^2)
           [,1]
[1,] 289320381
> (BLUE_var_mu <- sqrt(BLUE_est$Var_mu))
          [,1]
[1,] 2919.386
> (BLUE_var_var <- sqrt(BLUE_est$Var_sigma))
          [,1]
[1,] 3104.84
```

Gli stimatori BLUE di media e varianza per i Paesi poveri

```
> (BLUE_mean.poor <- BLUE_est.poor$muhat)
     [,1]
[1,] 3848
> (BLUE_sigma.poor <- BLUE_est.poor$sigmahat)
     [,1]
[1,] 1262
> (BLUE_var_mu.poor <- BLUE_est.poor$Var_mu)
      [,1]
[1,] 54505
> (BLUE_var_var.poor <- BLUE_est.poor$Var_sigma)
      [,1]
[1,] 37878
```

A sinistra l'output di R riguardante gli stimatori BLUE di media e varianza per la fascia di Paesi poveri. Nella terza e quarta riga sono presentate le deviazioni standard.

Gli stimatori MLE dei parametri shape e scale per i Paesi medio-ricchi

Per la stima dei parametri shape e scale è stato invece utilizzato il *metodo della massima verosimiglianza (MLE)*, in cui, oltre al campione, si presume anche la conoscenza della distribuzione della popolazione generatrice delle osservazioni campionarie.

```
mle_est_rich <- fitdistr(dati, "gamma", list(shape=3.19, scale=10105), lower=0.01)
mle_est_rich
   shape      scale
 3.24e+00   1.01e+04
(7.57e-01) (2.58e+03)
```

Ipotizzando dunque, in base ai precedenti risultati raggiunti con il test di Kolmogorov-Smirnov, che i dati campionari siano generati da una variabile Gamma si è proceduto alla stima MLE dei parametri, la quale, mediante il comando *fitdistr(),* richiede le osservazioni campionarie (*dati*), la popolazione generatrice dei

dati campionari (*Gamma*) e un valore iniziale dei parametri da stimare. Al di sotto degli stimatori MLE sono invece presentate le variazioni della stima.

Gli stimatori MLE dei parametri shape e scale per i Paesi poveri

Anche per i Paesi poveri, avendo seguito il medesimo procedimento appena presentato, vengono riportati gli stimatori MLE dei parametri shape e scale. Come per i Paesi ricchi si è assunta la provenienza dei dati campionari da una popolazione distribuita come una variabile casuale Gamma.

```
mle_est <- fitdistr(c(data), "gamma", list(shape=3.95, scale=990), lower = 0.1)
mle_est
 shape    scale
   3.2   1217.8
   1.6) ( 671.3)
```

Focus: gli stimatori bayesiani dei parametri shape e scale per i Paesi poveri

A conclusione di questa parte della trattazione riguardante la stima puntuale, si presenta ora un focus sulla determinazione dei parametri shape e scale nella popolazione dei Paesi poveri mediante un approccio di tipo Bayesiano, date le poche osservazioni in nostro possesso relative a tale fascia. Utilizzando i risultati ottenuti in precedenza dalla stima MLE, ipotizzeremo a priori una distribuzione normale di entrambi i parametri; come in precedenza, assumeremo inoltre che i dati provengano da una variabile casuale Gamma.

```
> stan_data
$N
[1] 7

$y
[1]  905 4823 2033 5137 2971 5137 6343

> stan.fit <-
+ "
+ data {
+     int<lower = 0> N;
+     real<lower = 0> y[N];
+ }
+ parameters {
+     real<lower = 0> shape;
+     real<lower = 0> scale;
+ }
+ model {
+     shape ~ normal(3.2, 1.6);
+     scale ~ normal(1217.8, 671.3);
+     y ~ gamma(shape, scale);
+ }
+ "
> set.seed(42)
> fit.gamma.norm <- stan_model(model_code = stan.fit)
```

Si riporta così alla sinistra il procedimento seguito per ottenere gli stimatori bayesiani dei parametri di nostro interesse, in cui nelle prime due righe vi sono la numerosità campionaria e i valori assunti nelle osservazioni, mentre nelle successive è riportato il codice utilizzato per generare gli stimatori bayesiani.

All'interno del codice si è quindi specificato:

-che la numerosità campionaria sia un numero intero e che i valori campionari appartengano all'insieme continuo

-che anche i parametri shape e scale siano appartenenti all'insieme continuo

-l'assunzione a priori riguardo la distribuzione dei parametri da stimare

-la forma della funzione di densità di probabilità nella popolazione (Gamma).

Tramite l'utilizzo di un seme casuale, necessario per poter generare le stime di nostro interesse, è stato così ottenuto il modello di stima bayesiana tramite il comando *stan_model().* In fondo alla pagina è presentato l'output di R riguardante i parametri shape e scale: si osserva come il valore ottenuto per il parametro shape (3,6) sia compreso tra quello dello stimatore MLE ricavato in precedenza (3,2) e il valore rilevato nel campione dei Paesi poveri utilizzato anch'esso nella stima di massima verosimiglianza (3,95); avendo inoltre solo 7 osservazioni campionarie, è possibile dedurre che le stime a priori abbiano influito maggiormente nella determinazione dei parametri. Per il parametro scale non si è invece ottenuto un risultato soddisfacente.

	mean	se_mean	sd	2.5%	25%	50%	75%	98%	n_eff	Rhat
shape	3.6	0.04	1.1	1.7	2.9	3.5	4.3	5.8	882	1
scale	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	934	1

Regressione

Si presenta ora un modello di regressione volto ad analizzare la dipendenza del Pil pro capite PPP dalle variabili presentate inizialmente, ovvero quelle "misure" dell'investimento in capitale umano da parte dei diversi Paesi del mondo. Verrà presa in considerazione solo la fascia dei Paesi medio-ricchi (da 7.384,27$ di Pil pro capite PPP in poi), ad esclusione di El Salvador (che con un Pil pro capite PPP di 7597,686$ rientrerebbe appena nella fascia oggetto dello studio, con valori notevolmente più bassi delle altre Nazioni) e del Lussemburgo (il cui reddito pro capite estremamente elevato pari a 103.722,9908$ si teme possa essere influenzato in prevalenza da fattori esterni alla nostra analisi).

Scelta delle variabili e Stepwise

La costruzione del modello è così iniziata dalla scelta di quelle variabili esplicative in grado di influenzare il Pil pro capite PPP, la proxy del tenore di vita esaminata.

```
Call:
lm(formula = gdp_cap ~ ., data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6151 -0.0859 -0.0017  0.0979  0.7946

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.25024    1.00511   11.19  1.2e-11 ***
exp           -0.02873    0.21651   -0.13   0.8954
articles.c     0.09128    0.07833    1.17   0.2541
researchers.c  0.00648    0.02561    0.25   0.8022
pupil.t       -0.51468    0.28461   -1.81   0.0817 .
research.exp   0.32503    0.11386    2.85   0.0082 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.26 on 27 degrees of freedom
Multiple R-squared:  0.768,     Adjusted R-squared:  0.725
F-statistic: 17.8 on 5 and 27 DF,  p-value: 8.21e-08
```

A fianco è presentato il modello con tutte le variabili prese in esame, le quali risultano essere: la spesa per istruzione come percentuale del Pil pro capite, gli articoli scientifici ogni 100.000 unità di forza lavoro, il numero di ricercatori nel settore R&D per milione di abitanti, il rapporto tra insegnanti e alunni nella scuola primaria e la spesa in R&D come percentuale del Pil. Come è possibile notare, in tale modello quasi tutte le variabili esplicative non sembrano essere significative, per cui si è deciso di procedere alla scelta di queste ultime tramite la Stepwise convenzionale; tutto ciò con l'obiettivo di ottenere un più efficiente modello di regressione.

Qui di fondo sono riportanti gli ultimi due passaggi del procedimento Stepwise (a sinistra) e le variabili ritenute più significative da questo (a destra), ovvero gli articoli scientifici ogni 100.000 unità di forza lavoro e il numero di ricercatori nel settore R&D per milione di abitanti. Tuttavia, anche tale modello non sembrava per noi esaustivo, per cui la scelta finale si è indirizzata verso un ulteriore e in parte differente modello di analisi.

```
Step:  AIC=-86
gdp_cap ~ articles.c + pupil.t + research.exp

               Df Sum of Sq  RSS   AIC
- articles.c    1     0.108 1.99 -86.6
<none>                     1.89 -86.5
+ researchers.c 1     0.004 1.88 -84.5
+ exp           1     0.001 1.88 -84.5
- pupil.t       1     0.375 2.26 -82.5
- research.exp  1     0.628 2.51 -79.0

Step:  AIC=-87
gdp_cap ~ pupil.t + research.exp

               Df Sum of Sq  RSS   AIC
<none>                     1.99 -86.6
+ articles.c    1     0.108 1.88 -86.5
+ researchers.c 1     0.016 1.98 -84.9
+ exp           1     0.001 1.99 -84.6
- pupil.t       1     0.750 2.74 -78.1
- research.exp  1     2.200 4.19 -64.1
```

```
Call:
lm(formula = gdp_cap ~ ., data = dati)

Residuals:
   Min     1Q Median     3Q    Max
-0.963 -0.214  0.019  0.254  0.908

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.4545     0.2239   37.76  < 2e-16 ***
articles.c        0.3559     0.0491    7.25  2.3e-10 ***
researchers.exp  -0.0376     0.0630   -0.60     0.55
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.36 on 80 degrees of freedom
Multiple R-squared:  0.637,     Adjusted R-squared:  0.628
F-statistic: 70.3 on 2 and 80 DF,  p-value: <2e-16
```

A sinistra gli ultimi due passaggi del procedimento Stepwise, a destra l'output al termine della Stepwise

La decisione finale è infine ricaduta sul modello presentato a destra, *con 57 osservazioni in comune per tutte le variabili,* le quali sono tutte espresse in logaritmi. È possibile notare, tramite il test-t effettuato sui parametri, come l'intercetta risulti significativamente diversa da zero, il coefficiente relativo alle pubblicazioni scientifiche e quello relativo al rapporto alunni-insegnanti appaiano significativi, mentre il coefficiente relativo al numero di ricercatori per milione di abitanti non risulti significativamente diverso da zero.

```
Call:
lm(formula = gdp_cap ~ ., data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9167 -0.2311  0.0046  0.2091  0.7371

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.9560     0.6680   14.90  < 2e-16 ***
articles.c     0.3487     0.0477    7.31  1.4e-09 ***
researchers.c  0.0123     0.0182    0.67    0.5034
pupil.t       -0.5590     0.1945   -2.87    0.0058 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.33 on 53 degrees of freedom
Multiple R-squared:  0.689,     Adjusted R-squared:  0.671
F-statistic: 39.2 on 3 and 53 DF,  p-value: 1.77e-13
```

Il segno dei parametri indica come sia l'incremento della qualità del sistema primario di istruzione (rappresentato dalla voce *pupil.t*, la quale assume valori maggiori in corrispondenza di bassi standard qualitativi) sia il maggior "impegno" dei diversi Stati nella pubblicazione scientifica e nello sviluppo di nuove conoscenze (indicato dalla voce *articles.c*) concorrano positivamente alla dinamica del Pil pro capite PPP. I ricercatori ogni 100.000 abitanti non hanno invece un effetto significativo sulla variabile risposta, il loro contributo è praticamente nullo; tale variabile potrebbe essere rimossa dal modello, tuttavia, poiché una sua rimozione avrebbe comportato una peggiore distribuzione grafica dei residui, si è deciso in ultima istanza di tenerla.

L'indice R^2 modificato, che misura la bontà di adattamento del modello al netto del numero di regressori considerati, è pari a 0,671, il modello spiega così una parte rilevante della variabilità della variabile risposta.

Validazione del modello e analisi dei residui

1) Verifica che i residui siano a media nulla

Il primo passo per validare il nostro modello è stato quello di verificare che $E(\varepsilon) = 0$, ovvero che gli errori fossero a media nulla.

```
> t.test(modello$residuals)

        One Sample t-test

data:  modello$residuals
t = -1e-16, df = 56, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.086  0.086
sample estimates:
mean of x
 -4.3e-18
```

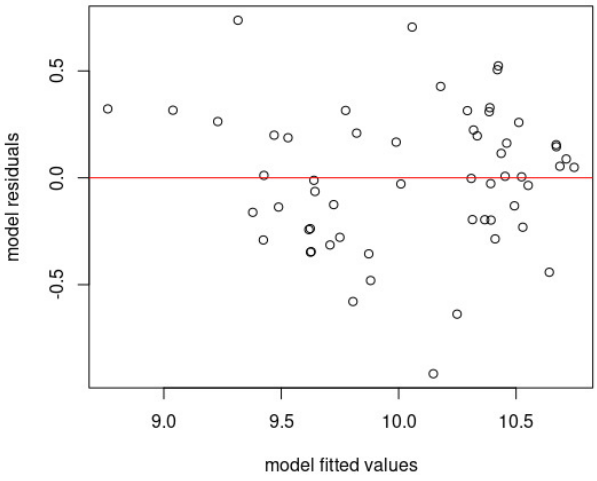
Tale verifica è stata condotta tramite il t-test, il quale prevede come schema di ipotesi:

- $H_0: E(\varepsilon) = 0$
- $H_1: E(\varepsilon) \neq 0$

Essendo il p-value elevato (pari ad 1) non è possibile rifiutare l'ipotesi nulla, la media dei residui può essere considerata pari a zero.

Contestualmente al t-test è possibile effettuare una verifica grafica di tale assunzione, constatando visivamente che i residui si distribuiscano casualmente intorno allo zero.

Il grafico rappresentante la distribuzione dei residui, questi risultano distribuirsi casualmente intorno allo zero



2) Verifica dell'omoschedasticità dei residui

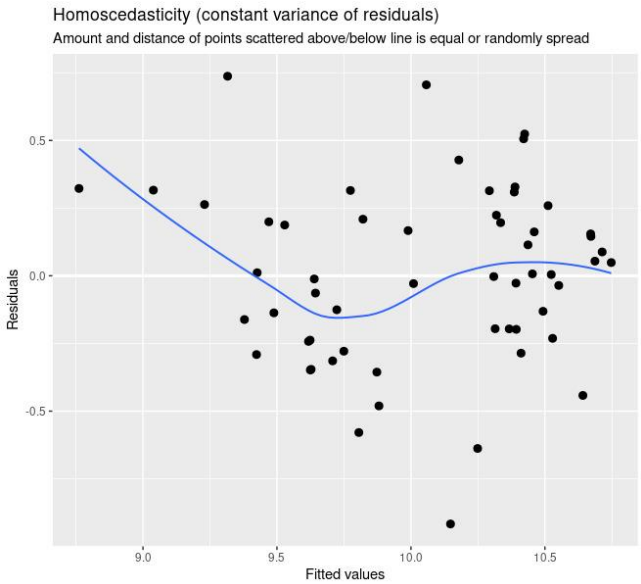
Per far sì che siano rispettate le ipotesi deboli dello schema di regressione, si è reso successivamente necessario verificare la omoschedasticità dei residui, ovvero che questi avessero varianza costante.

La verifica in tale caso è stata condotta mediante il test di Breusch-Pagan, il cui schema di ipotesi risulta essere così formulato:

- $H_0: Var(\varepsilon) = \sigma^2 I$
- $H_1: altrimenti$

```
> bptest(modello)

        studentized Breusch-Pagan test

data:  modello
BP = 5, df = 3, p-value = 0.2
```

Essendo il p-value anche in questo caso maggiore di 0,05, si accetta l'ipotesi di omoschedasticità dei residui.



Dal grafico è possibile notare come non si evidenzino strutture nei dati, gli errori hanno varianza costante

3) Verifica dell'incorrelazione tra i residui

Vi è un'ultima ipotesi da validare per assicurare che il nostro modello rispetti quantomeno le ipotesi deboli del teorema di Gauss-Markov: i residui devono essere tra loro incorrelati.

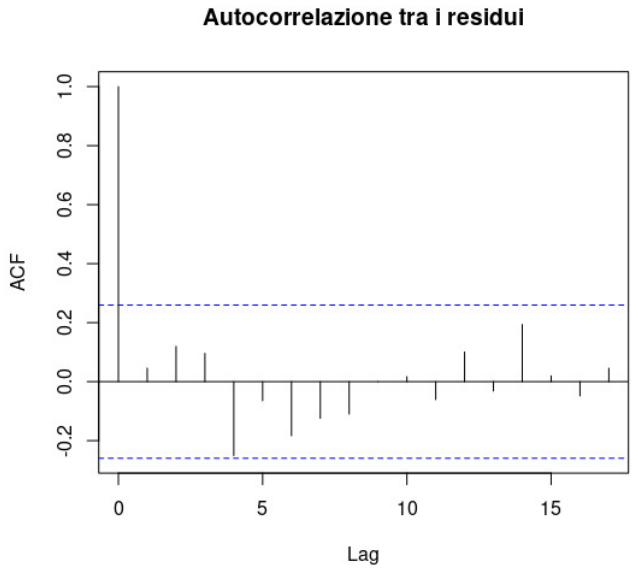
Per tale verifica di è utilizzato il test di Durbin-Watson, il cui schema di ipotesi risulta essere così formulato:

- $H_0: Corr(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i,j}$
- $H_1: Corr(\varepsilon_i, \varepsilon_j) \neq 0$

```
        Durbin-Watson test

data:  formula(modello)
DW = 2, p-value = 0.4
alternative hypothesis: true autocorrelation is greater than 0
```

L'ipotesi nulla di incorrelazione tra i residui è accettata in base al valore pari a 0,4 assunto dal p-value.

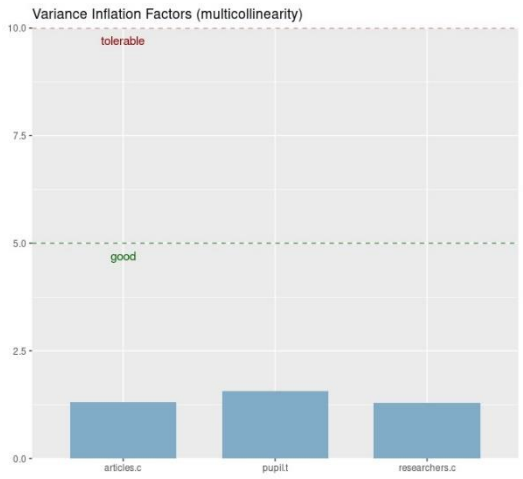
Graficamente tale incorrelazione è visibile tramite il correlogramma: non evidenziandosi valori esterni all'intervallo definito dalle linee tratteggiate si esclude l'ipotesi di strutture autocorrelative dei residui; anche il fatto che le "barre" si alternino al di sotto ed al di sopra dello zero conferma tale assunzione.



Al termine di tale verifica è possibile così concludere che il modello rispetta le ipotesi deboli del teorema di Gauss-Markov.

Verifica dell'assenza di multicollinearità

Per essere attendibile, il modello richiede che sia confermata l'assenza di multicollinearità, la quale si presenta nei casi in cui le variabili esplicative siano tra loro notevolmente correlate. La presenza di multicollinearità potrebbe portare a valori anomali dei parametri stimati: per tale motivo è stata effettuata, tramite il *Variance Inflaction Factor (VIF),* la verifica su tale effetto di disturbo.




```
vif(modello)
 articles.c researchers.c       pupil.t
        1.3           1.3           1.6
```

Il valore minore che il VIF può assumere è 1, mentre risulta non accettabile tra 5 e 10. Avendo assunto valori molto bassi in corrispondenza delle variabili indipendenti del modello, è possibile concludere che queste siano tra loro incorrelate.

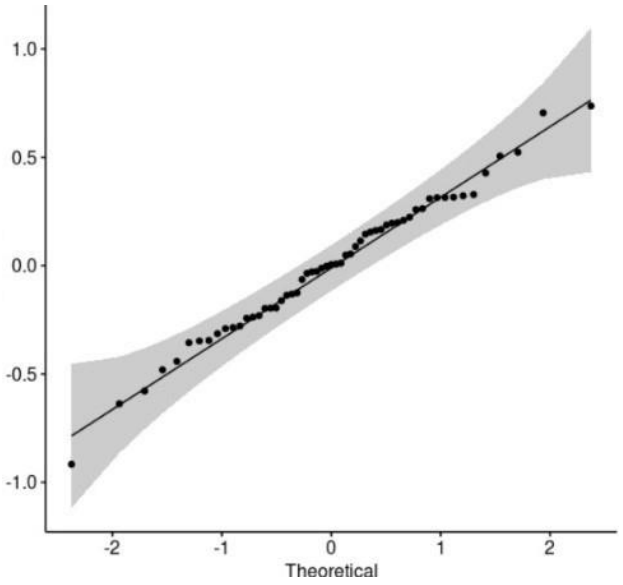
Verifica della normalità dei residui

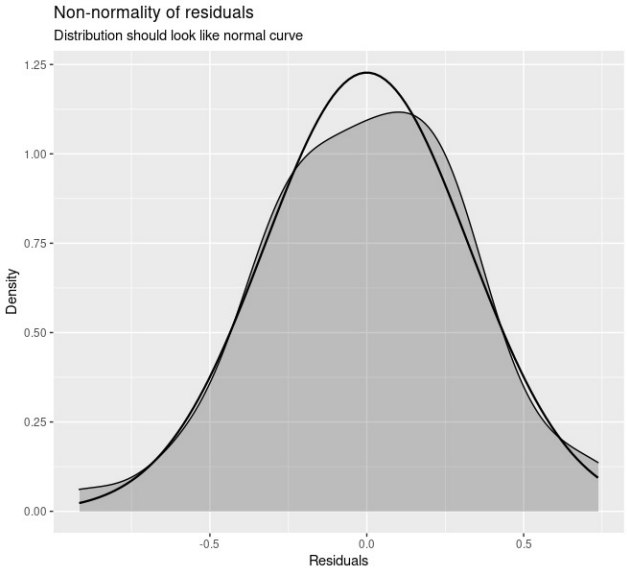
Occorre infine valutare se i residui possano essere assunti come normali (c.d. ipotesi forti del teorema di Gauss-Markov). A tal fine è stato impiegato il test di Shapiro-Wilk, il cui sistema di ipotesi è dato da:

- H_0: *i residui si distribuiscono normalmente*
- H_1: *i residui non si distribuiscono normalmente*

```
> shapiro.test(modello$residuals)

        Shapiro-Wilk normality test

data:  modello$residuals
W = 1, p-value = 0.9
```

Dato un valore del p-value di 0,9, viene accettata l'ipotesi nulla di normalità dei residui. Tale verifica può avvenire, come in precedenza, a livello grafico: vengono pertanto riportati in fondo il qq-plot e un grafico che mette a confronto la distribuzione dei residui con quello di una normale standard.



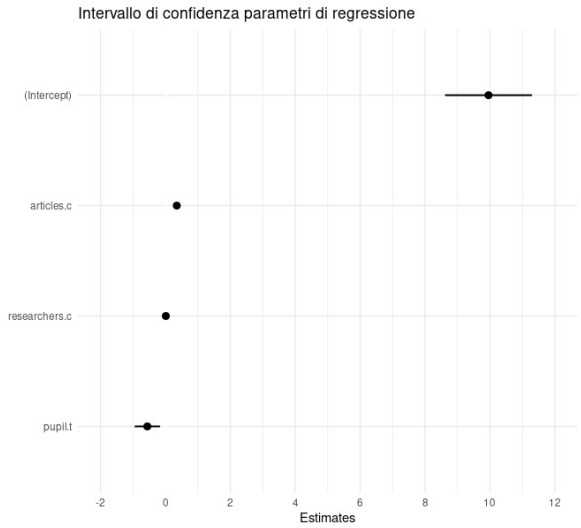


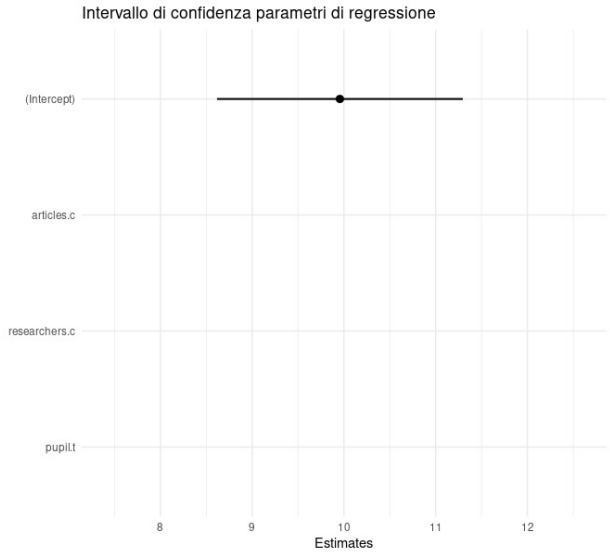
Il rispetto delle ipotesi forti comporta che anche la distribuzione della variabile dipendente Y possa essere assunta normale, essendo questa una combinazione lineare di v.c. normali. Le medesime considerazioni possono essere fatte per i parametri del modello, essendo questi a loro volta combinazioni lineari della variabile Y. Ciò consente di stimare gli intervalli di confidenza per tali parametri, gli intervalli di predizione e di attuare il test ANOVA per testare la significatività del modello.

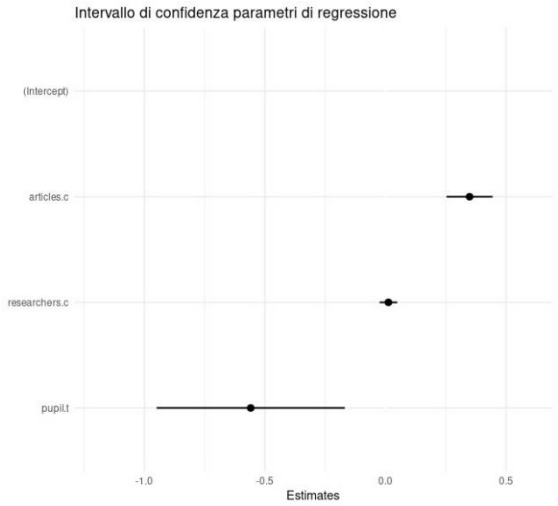
Intervalli di confidenza

Di fianco sono riportati gli intervalli di confidenza (con $\alpha = 0{,}05$) dei parametri del modello di regressione. La dipendenza diretta tra Pil pro capite PPP e pubblicazioni scientifiche ogni 100.000 unità di FL viene confermata, come anche quella indiretta tra la variabile risposta ed il rapporto alunni-insegnanti nella scuola primaria. Il segno del parametro relativo ai ricercatori per milione di abitanti risulta invece incerto a tale livello di confidenza, non discostandosi significativamente da zero.

	2.5 %	97.5 %
(Intercept)	8.616	11.296
articles.c	0.253	0.444
researchers.c	-0.024	0.049
pupil.t	-0.949	-0.169



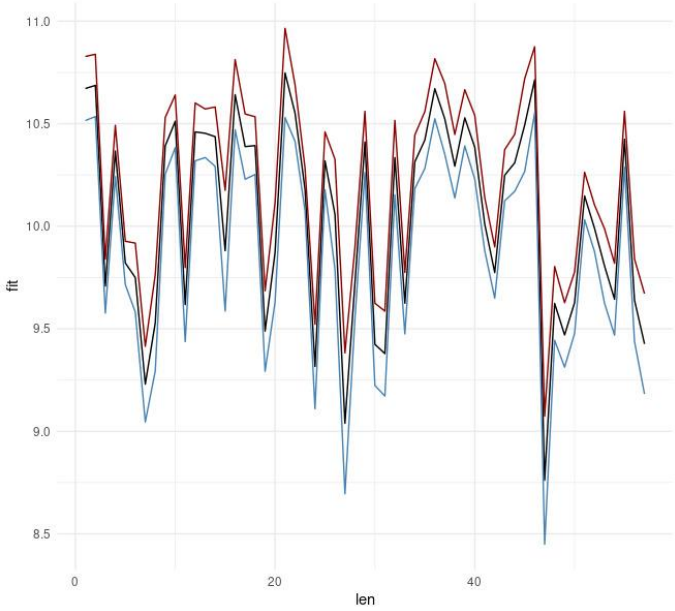




In alto a sinistra gli intervalli di confidenza dei parametri. In alto a destra uno zoom sull'intervallo di confidenza dell'intercetta. A sinistra uno zoom sugli intervalli di confidenza dei parametri associati alle variabili esplicative.

Intervallo di previsione

Dopo aver visto gli intervalli di confidenza dei parametri del modello si presenta ora l'intervallo di previsione, in modo da poter per l'appunto cercare di "prevedere" valori della variabile risposta Y in corrispondenza di valori non osservati delle variabili esplicative per ciascuna delle diverse osservazioni. Di fianco è riportata la rappresentazione grafica di tale intervallo, dove la linea nera è il valore predetto mentre le linee rossa e blu rappresentano rispettivamente l'estremo superiore ed inferiore dell'intervallo di confidenza.



La significatività globale del modello: l'ANOVA sullo schema di regressione

Al termine dell'analisi riguardante il modello di regressione, si intende ora verificare la significatività del modello nella sua interezza, ossia si vuole testare che la devianza spiegata dal modello sia maggiore di quella di un modello "vincolato" con la presenza della sola intercetta.

```
Analysis of Variance Table

Response: gdp_cap
              Df Sum Sq Mean Sq F value  Pr(>F)
articles.c     1  11.73   11.73  105.08 3.5e-14 ***
researchers.c  1   0.46    0.46    4.13  0.0472 *
pupil.t        1   0.92    0.92    8.26  0.0058 **
Residuals     53   5.92    0.11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'ANOVA, tramite un test-F, mette così a confronto il modello da noi utilizzato con un modello con la sola β_0. Dati i bassi livelli del p-value (minori di 0,05) si rifiuta l'ipotesi che la devianza spiegata dal modello sia pari a zero, evidenziando una significatività complessiva dei regressori utilizzati.

Conclusione

A conclusione dello studio, le osservazioni che si possono fare in relazione al rapporto tra il tenore di vita dei diversi Paesi e l'investimento in capitale umano promosso da quest'ultimi appare raffigurarne una relazione diretta. Sebbene nessuna delle voci di spesa in senso proprio abbia contribuito al modello, gli indici di qualità dell'istruzione a livello di scuola primaria e lo sviluppo dello stock di conoscenze nel settore della ricerca risultano comunque essere correlati positivamente ai livelli di reddito pro capite delle differenti Nazioni.

Tra le variabili esplicative del nostro modello di regressione è stato soprattutto il rapporto alunni-insegnanti ad essere risultato interessante; difatti, pur avendo analizzato in tale contesto unicamente una fascia di Paesi da noi definiti medio-ricchi, l'incidenza di tale misura non sembra irrilevante: un buon livello di qualità dell'istruzione primaria si riflette nel tenore di vita nei vari Paesi. La ricerca ed il progresso scientifico concorrono anch'essi a "spiegare" la diversa distribuzione del reddito pro capite, anche se ciò è apparso con meno evidenza dal nostro studio.

L'investimento nell'individuo rimane, in definitiva, un elemento importante tra tutte le differenti tipologie di investimenti atte a favorire il progresso delle Nazioni a livello mondiale; al fianco di quello che nell'introduzione è stato definito "capitale fisico" si colloca così questo "capitale umano", forse a volte tenuto solo marginalmente in considerazione, ma non per questo di minore interesse.