Assignment

Adnan Sardi

June 2022

1 Obiettivo

L'obiettivo dell'assignment è quello di effettuare un analisi dei dati libera sul data set "EU_Econ_Data_2021.csv". Dopo aver scelto la variabile dipendente e quelle indipendenti, effettuare una regressione lineare multipla.

2 Analisi esplorativa dei dati

Dobbiamo innanzitutto scaricare e leggere il data set "EU_Econ_Data_2021.csv" su R Studio, per tali operazioni sono stati utilizzati i seguenti comandi:

- setwd("/Users/adnan/Desktop/R/Materiale Big Data")
- Data_set_eu ← read.csv("EU_Econ_Data_2021.csv", sep=";")

2.1 Normalizzazione dataset

Andiamo ora a normalizzare il nostro dataset poichè le nostre osservazioni hanno scale di valore molto differente.
Per fare ciò utilizziamo la funzione *sapply*.

- data_set_eu_norm ← sapply(Data_set_eu[2:5], scale)
- head(data_set_eu_norm)
- dati_norm ← as.data.frame(data_set_eu_norm)

Qui di seguito possiamo visualizzare i nostri dati normalizzati.
Dai dati in possesso possiamo già intuire come la variabile dipendente y sia la "Per_capita_GDP" mentre "Capital_perc_GDP", "Unemployment_rate" e "Inflation_rate" sono le nostre variabili indipententi $x_{1,2,3}$.

dati_norm

Filter

	Per_capita_GDP	Capital_perc_GDP	Unemployment_rate	Inflation_rate
1	0.41083949	-0.5450434	-0.11644810	0.30695248
2	-1.14854028	-0.2694597	-0.46195346	-0.02790477
3	-0.54799725	0.2265911	-1.32571684	0.39066680
4	1.17822974	-0.6001602	-0.53105453	-0.78133359
5	0.38089255	-0.6552769	-1.04931256	0.30695248
6	-0.63676853	4.4154643	-0.15099864	1.39523856
7	2.28626639	-0.7655104	-0.15099864	-0.36276202
8	-0.56564455	0.5572916	2.78579688	-1.86961966
9	-0.24906265	-0.1592262	2.82034741	0.13952386
10	0.23971414	-0.3796932	0.43636047	-0.61390496
11	-0.78650321	0.6124084	0.33270886	-0.11161908
12	-0.07847207	-0.7103937	0.98916903	-0.78133359
13	-0.17366054	-0.2143429	0.29815832	-0.44647634
14	-0.82179781	0.1163576	0.33270886	0.30695248
15	-0.72072690	0.1714743	0.15995618	1.47895287
16	3.12210678	-0.6552769	-0.46195346	0.55809542
17	-0.77580787	0.7226419	-0.87655988	1.98123875
18	-0.32820813	-0.1041094	-1.08386309	-1.78590535
19	0.73811672	-0.6001602	-0.84200935	-0.02790477
20	0.46271186	-0.6552769	-0.15099864	-0.02790477
21	-0.78008601	0.2265911	-1.11841363	1.98123875
22	-0.54799725	0.3368246	-0.01279649	-1.61847672
23	-1.00468803	0.5572916	-0.35830185	1.06038130
24	-0.36938516	-0.6001602	-0.63470613	-0.69761928
25	-0.66885453	0.1163576	0.05630458	-0.02790477
26	0.49372833	-0.6001602	0.36725940	-0.61390496
27	0.89159477	-0.5450434	0.74731529	-0.11161908

Figure 1:

2.2 Analisi esplorativa tramite grafici

Per farci un idea di come si distribuisce la nostra variabile dipendente eseguiamo un istogramma.

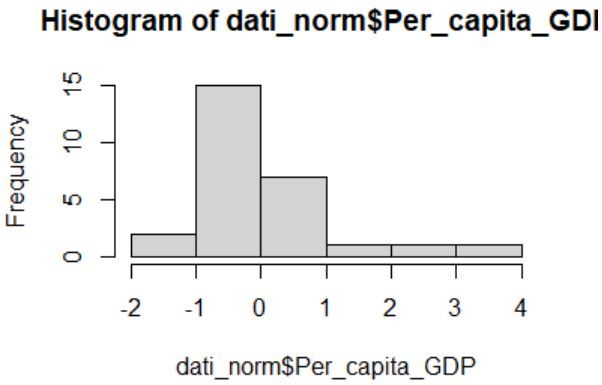

Figure 2:

I dati nonostante siano stati normalizzati si distribuiscono in un intervallo di valori compreso tra -2 e +4, questo potrebbe portare in futuro a problematiche di vario genere in un analisi più dettagliata. Inoltre i dati sembrano accostarsi ad una distribuzione normale anche se poco accentuata.

Possiamo effetturare degli scatter-plot tra la variabile dipendente y e le varie varibili indipendenti x-esime.

Da questi tre plot possiamo iniziare a fare qualche deduzione che poi possono essere confermate o rifiutate da vari test statistici come la correlazione.

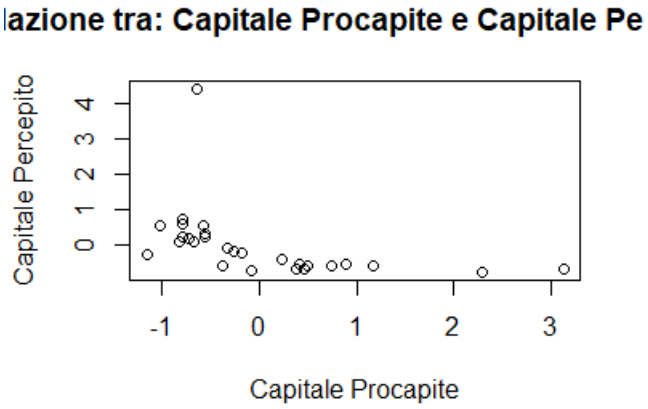

Figure 3:

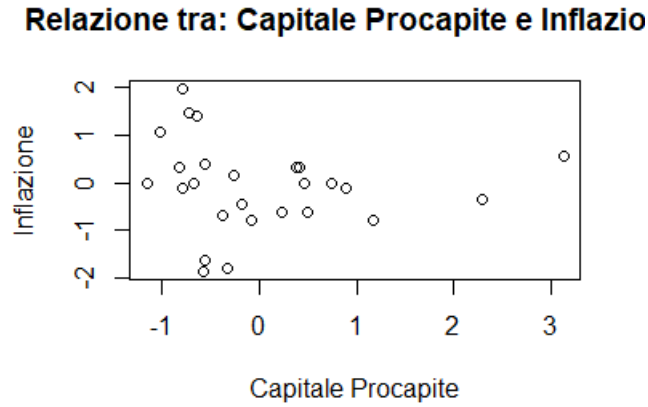

Figure 4:

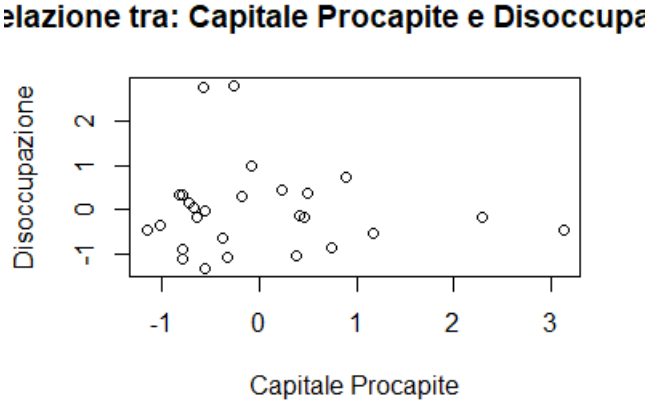

Figure 5:

La relazione tra il Capitale Procapite con l'inflazione e la disoccupazione non sembrerebbe darci informazioni rilevanti essendo i punti distribuiti senza nessuna tendenza di fondo. Mentre per il primo plot possiamo notare come per un Capitale percepito basso abbiamo dei valori di Capitale Procapite addensato in valori compresi tra -1 e 0, ciò ci suggerisce l'esistenza di una correlazione tra queste due variabili.

2.2.1 Correlazione

Per poter calcolare la correlazione tra variabili e plottarle tramite grafico in R Studio dobbiamo prima scaricare la libreria "corrplot" tramite la seguente operazione:

- library("corrplot")

Le seguenti righe di codice servono per estrapolare due osservazioni dal nostro dataset per poi crearne uno soltanto tramite la funzione *cbind* ed infine utilizziamo la funzione *cor* per determinare la correlazione tra le due osservazioni estrapolate.
Qui di seguito riportiamo le righe di codice appena citate:

- CPC_CP ← cbind(dati_norm$Per_capita_GDP, dati_norm$Capital_perc_GDP)
- corr_CPC_CP ← cor(CPC_CP)

Con CPC_CP intendiamo Capitale Procapite correlato con Capitale Percepito mentre con CPC_U intendiamo Capitale Procapite correlato con Dissocupazione mentre CPC_I intendiamo Capitale Procapite correlato con Inflazione.
Riportiamo qui di seguito i valori trovati:

$$corr(CPC\_CP) = -0,439 \tag{1}$$

$$corr(CPC\_U) = -0,070 \tag{2}$$

$$corr(CPC\_I) = -0,149 \tag{3}$$

Abbiamo una conferma di come l'inflazione e la disoccupazione non sono osservazioni correlate al aumento o meno del capitale procapite di un paese. Ciò non si può dire sulla variabile capitale percepito poichè il valore di correlazione ottenuto è circa 0,5. Un informazione importante la si può ottenere dal segno (negativo) che ci aiuta a capire come le due osservazioni sembrerebbero una inversamente proporzionale all'altra.
I valori di correlazione possono essere plottati tramite grafico utilizzando la funzione *corrplot*. Riportiamo la sintassi della prima funzione con il rispettivo grafico:

```
corrplot(cor(CPC_CP),
method = "shade",
type = "full",
diag = TRUE,
tl.col = "black",
bg = "white",
title = "",
col = NULL)
```

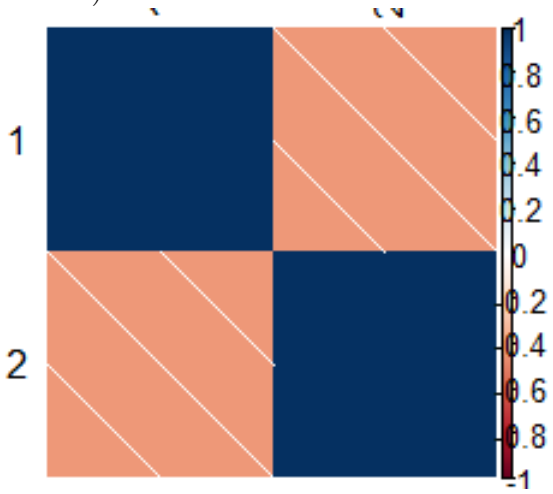

Figure 6:

Il valore della correlazione può essere letto tramite la scala di colori che si trovano sulla destra dell'immagine. Ovviamente il primo e il quarto quadrante sono di colore blu che corrisponde ad una corrleazione uguale ad uno, se avessimo inserito la stessa immagine per le altre due correlazioni calcolate avremmo avuto il secondo e il terzo quadrante di colore bianco. Sul file R allegato a questo è possibile visualizzarli.

3 Ricerca del miglior modello

Eseguita un'analisi esplorativa passiamo alla ricerca del miglior modello. Inizialmente andiamo a creare il modello nullo e quello saturo tramite la funzione *lm*, dove con modello nullo si intende il modello con il solo valore di intercetta mentre con modello saturo il modello con tutti i parametri indipendenti. La sintassi su R è la seguente:

- intercept_only_model $\leftarrow$ lm(Per_capita_GDP $\sim$ 1,data = dati_norm)
- full_model $\leftarrow$ lm(Per_capita_GDP $\sim$., data = dati_norm)

Ora possiamo utilizzare la funzione *summary* applicata ai nostri due modelli per poter fare qualche considerazione più dettagliata.
Il plot è il seguente:

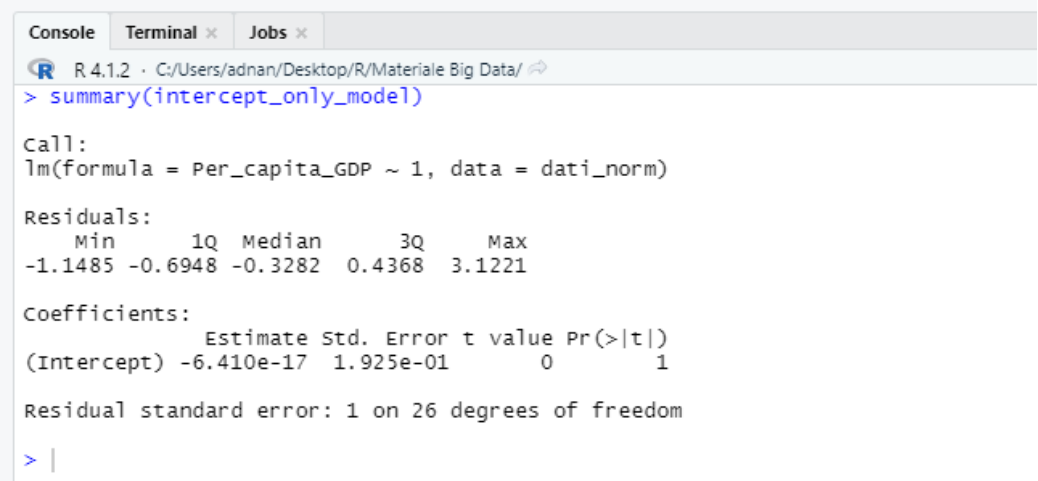

Figure 7: Caption

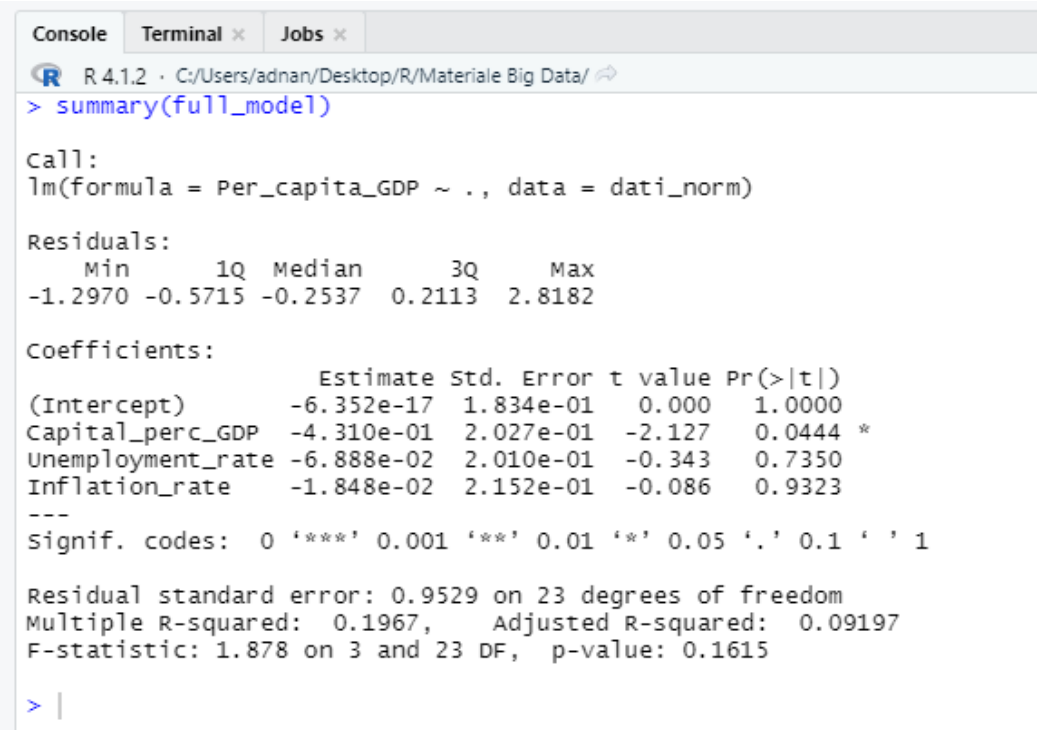

Figure 8: Caption

Il modello nullo non ci fornisce informazioni rilevanti mentre quello saturo stima il parametro Capital_perc_GDP come parametro rilevante con un valore di t minore di quello significativo di riferimento(0.05). Possiamo ora calcolare il miglior modello tramite l'algoritmo step-wise sia partendo dal modello nullo che da quello saturo, la funzione utilizzata è *step*.
Qui di seguito sono riportati i tre modelli utilizzati in R.

- stepwise_forward ← step(intercept_only_model, direction = 'forward', scope = formula(full_model), trace = 1)
- stepwise_backward ← step(full_model, direction = 'backward', scope = formula(full_model), trace = 1)
- stepwise_both ← step(intercept_only_model, direction = 'both', scope = formula(full_model), trace = 1)

Su R studio è possibile visualizzare la summary di tutte e tre le funzioni, per semplicità qui viene riportata quella della sola funzione *stepwise_forward*.

```
> summary(stepwise_forward)

Call:
lm(formula = Per_capita_GDP ~ Capital_perc_GDP, data = dati_norm)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2668 -0.5678 -0.3211  0.2028  2.8346

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6.410e-17  1.764e-01   0.000    1.000
Capital_perc_GDP -4.388e-01  1.797e-01  -2.442    0.022 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9164 on 25 degrees of freedom
Multiple R-squared:  0.1926,    Adjusted R-squared:  0.1603
F-statistic: 5.963 on 1 and 25 DF,  p-value: 0.02203

> |
```

Figure 9: Caption

Sembra che il nostro modello con la variabile Per_capita_GDP come variabile dipendente viene influenzata dalla variabile Capital_perc_GDP più che dalle altre, questo fatto non ci sorprende essendo che tale risultato poteva essere dedotto dall'analisi esplorativa eseguita in precedenza.
Inoltre il valore di p è inferiore a quello standard di 0.05 e questo ci suggerisce di rifiutare l'ipotesi nulla, cioè assumere beta uguale a zero. Il valore di beta 1 è leggermente inferiore a zero, ciò significa che un aumento di CP diminuisce

la nostra variabile dipendente, questo risultato era stato ipotizzato tramite una ricerca di correlazione tra le due variabili. L R^2 invece è basso mentre il valore di F test è uguale a 5.96.

4 Cross Validation

Per poter stimare la bontà del nostro modello abbiamo più indici e tecniche da poter utilizzare. In questo assignment ci limiteremo a utilizzare la tecnica "Leave-one-out Cross Validation". Per prima cosa abbiamo installato il pacchetto contenete le istruzioni per applicarla. Di seguito riportiamo lo script eseguito su R per poi commentarlo assieme ai risultati ottenuti.

- install.packages("caret")
- library(caret)
- LooCV ← trainControl(method = "LOOCV")
- modello ← train(Per_capita_GDP Capital_perc_GDP , data = dati_norm, method = "lm", trControl = LooCV)
- print(modello)

```
> print(modello)
Linear Regression

27 samples
 1 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 26, 26, 26, 26, 26, 26, ...
Resampling results:

  RMSE      Rsquared    MAE
  1.474549  0.03994492  0.8659207

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

Figure 10: Caption

Abbiamo 3 indici significativi. RMSE, esso ci indica la differenza che intercorre tra i valori osservati e quelli predetti, nel nostro caso assume un valore di 1.47. Il secondo indice è RSQUARED che indica la correlazione tra i valori osservati e predetti, nel nostro caso è molto basso tendente allo zero. L'ultimo indice MAE indica la media dell'errore assoluto che in questo caso vale 0.87.

5 Distribuzione errori

Un importante analisi da effettuare sul nostro modello è quello sugli errori. Ci sono moltissimi test che si possono eseguire, in questo assignment riportiamo semplicemente il grafico QQplot senza la funzione utilizzata in R.

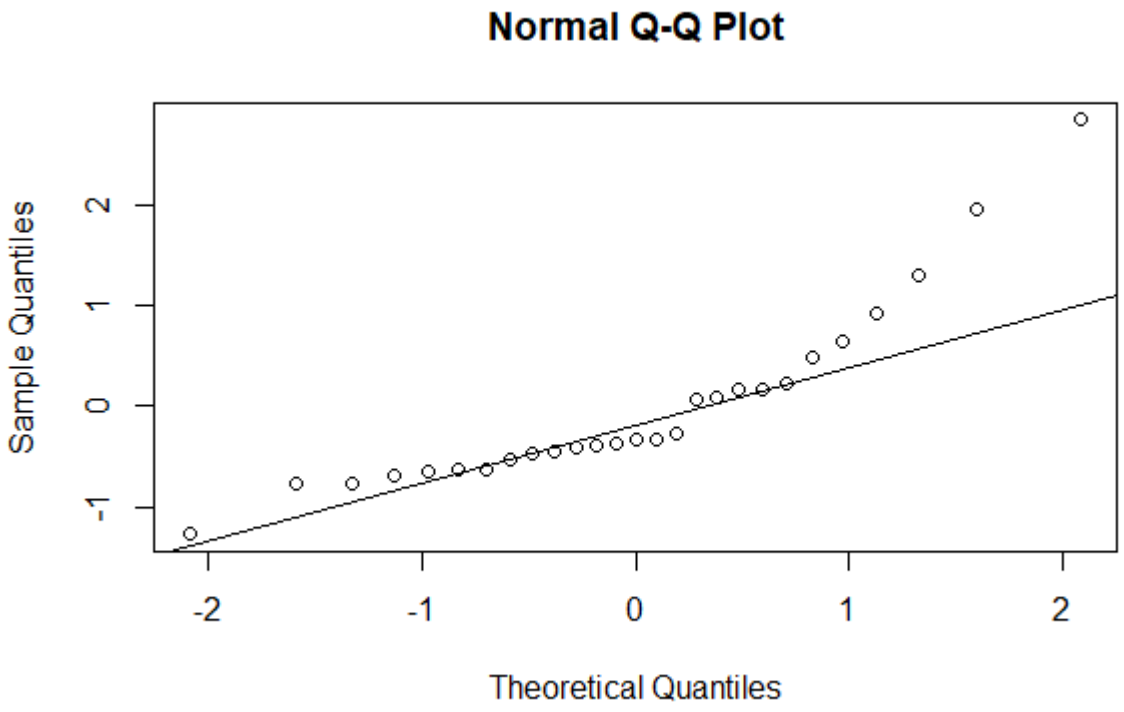

Figure 11: Caption

Si vede facilmente come i nostri errori seguono un andamento lineare fino ad un valore di 1 sull'asse delle ascisse per poi allontanarsi sempre più. Un test che può essere eseguito è quello do Jarque_bera per verificare la normalità degli errori.

6 Plot risultati finali

Plottiamo i risultati finali ottenuti utilizzando la funzione *ggplot* aggiungendo la funzionalità *geom_smooth* che aggiunge una retta di regressione con un intervallo di errore.

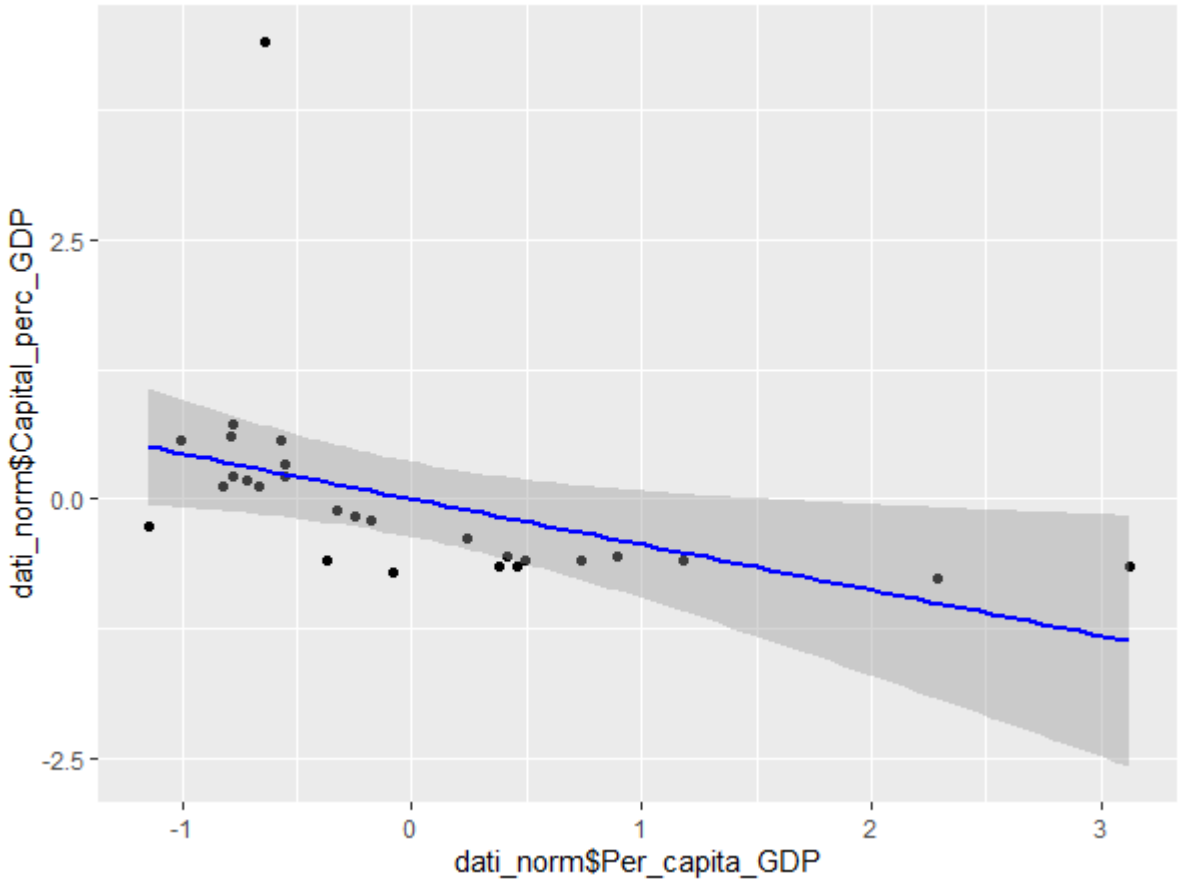

Figure 12: Caption

Come da ipotesi la banda di errore si restringe su valori che partono da -0.45 fino a 0 per poi divergere all'aumentare di "Per_capita_GDP". Di seguito sono riportati le funzioni utilizzate su R.

- Per_capita_GDP.graph←ggplot(dati_norm, aes(x=dati_norm$Per_capita_GDP, y=dati_norm$Capital_perc_GDP))+ geom_point()
- Per_capita_GDP.graph ← Per_capita_GDP.graph + geom_smooth(method = "lm", col = "blue")
- Per_capita_GDP.graph

7 Conclusioni

Per verificare l'effettiva bontà del nostro modello bisognerebbe effetturare altri test diagnostici, sicuramente possiamo affermare come il capitale percepito influisce sul capitale procapite di un paese anche se in modo inversamente proporzionale. Per un miglioramento del modello sarebbe opportuno possedere più unità statistiche poichè averne 27 è molto limitante ai fini statistici. L'analisi effettuata non è sicuramente quella più ottimale, miglioramenti che possono essere condotti è quello dell'utilizzo dell'algoritmo delle k-means, verificare normalità dei residui, utilizzare funzioni più complesse ma più precise in ambito regressioni lineari come *leaps* oppure eseguire una pulizia iniziali dei dati affinchè non vadano ad influenzare negativamente sui risultati fnali.