Progetto Modelli Statistici: un'analisi del tasso di turisticità

Enrico Guerriero
15 febbraio 2023

1 Introduzione

Il progetto che segue è un'analisi del tasso di turisticità. Come prima cosa, si fornisce la definizione di tale tasso:

$$\text{Tasso di turisticità} = \frac{\text{Giornate di presenza dei turisti}}{\text{Popolazione residente}}$$

L'obiettivo dell'analisi è studiare la variazione del tasso nel tempo e nello spazio. Più nello specifico, l'analisi temporale riguarda l'intervallo 2000 - 2020, e prende in esame i gruppi di regioni d'Italia:

I cinque gruppi, come si può intuire, prendono il nome di Nordest, Nordovest, Centro, Sud e Isole.

L'altra analisi che si vuole effettuare è un'analisi trasversale: si fissano 4 momenti (2000 - 2010 - 2019 - 2020) nel tempo e si osserva il comportamento del tasso rispetto alle variabili esplicative. Questo tipo di analisi, invece, avviene con la divisione regionale tradizionale.

I dati sul tasso di turisticità sono stati scaricati dal sito della Banca d'Italia, nella sezione del turismo internazionale.

Le variabili esplicative

- Dati sul traffico aereo: sono stati presi dal sito di Asseaeroporti. Le variabili effettivamente utilizzate nel modello sono:
  - **Movimenti aerei**: Numero totale degli aeromobili in arrivo/partenza.
  - **Passeggeri**: Numero totale dei passeggeri in arrivo/partenza, inclusi i transiti diretti (ossia i passeggeri che transitano in un aeroporto e ripartono utilizzando un aeromobile con lo stesso numero di volo dell'arrivo).

  I dati sono stati scaricati da un totale di 274 database mensili, contenenti altre variabili che in fase di pulizia o di valutazione dei modelli sono state scartate. In una fase di download, sono stati creati 2 database 276×39, uno per ciascuna variabile, in cui ciascuna colonna corrisponde ad un aeroporto e ciascuna riga ad un mese.
  In seguito avviene la selezione e l'aggregazione dei dati: i dati sono stati aggregati annualmente e sono stati

scartati i dati relativi al 2021 e al 2022, disponendo della sola serie storica annuale e compresa tra il 2000 e il 2020 della variabile risposta.

- Dati sul turismo: sono stati presi dalla sezione sul turismo della Banca d'Italia, come la variabile risposta, e dal sito dell'Istat. Le variabili esplicative ricavate da questi due database sono:
  - **Valore aggiunto del turismo**: valore aggiunto ai prezzi base nel settore del turismo, espresso in milioni di euro ai prezzi correnti.
  - **ULA turismo**: Unità di lavoro nel settore del turismo, medie annue espresse in migliaia di euro.
  - **Produttività del lavoro nel turismo**: Valore aggiunto del settore del turismo per ULA dello stesso settore, espresso in migliaia di euro concatenati
  - **Spesa per regione di destinazione**: espressa in milioni di euro
  - **Spesa per motivo**: questa variabile è l'unica che non effettua una divisione delle determinazioni per regione, ma invece dà una proporzione, a partire dalla spesa totale dei turisti sul territorio italiano, della spesa effettuata per viaggi di lavoro o per motivi personali (tra i quali è specificata la cifra nel caso in cui il motivo sia la semplice vacanza)

  Le variabili sopracitate sono state scaricate da file excel in cui a ciascuna variabile era dedicata una pagina, quindi la pulizia è stata più semplice. Anche in questo caso, alla fine per ciascuna variabile si ha un dataframe 21×20 con come righe gli anni dal 2000 al 2021 e come colonne le regioni.
- Dati sulle regioni: i dati sono stati presi da database Istat e servono per assegnare alle regioni dei valori che prescindono dal turismo ma possono essere influenti nella determinazione del tasso:
  - **PIL**: il prodotto interno lordo è proposto come variabile esplicativa nel modello sebbene il PIL sia in parte spiegato dal turismo; infatti il turismo in Italia, mediamente, spiega circa l'1.3% del PIL. Pertanto questa dipendenza, ad eccezione di alcune regioni, può essere trascurata in favore di un'interpretazione del PIL come ricchezza del territorio.
  - **Popolazione**: la popolazione nelle varie regioni è il denominatore della variabile risposta che si va a studiare, pertanto la correlazione tra le due variabili è matematica. Tuttavia, la popolazione può essere utile per dare una dimensione alle regioni, dal momento che tutti i dati a disposizione sono assoluti mentre il tasso è pro capite.

  Anche per queste due variabili sono stati creati i due database di rito 21×20.
- Dati sui beni culturali: i dati sono stati presi da wikipedia, non avendo trovato dei dataset già presenti online, e incollati in un file csv (fortunatamente sono davvero pochi dati):
  - **Bandiere blu**: Numero di spiagge con bandiera blu in ciascuna regione
  - **Parchi nazionali**: Numero di parchi nazionali in ciascuna regione (sia numero di parchi esclusivamente in una regione, sia condivisi)
  - **Siti UNESCO**:Numero di siti culturali e ambientali considerati Patrimonio dell'Umanità dall'UNESCO in ciascuna regione

  Da queste variabili si ricava un unico dataframe 4×20 in cui le colonne sono le regioni e le righe le 4 variabili sopra elencate. Questo è l'unico dataframe che non è in serie storica, ed infatti verrà utilizzato nella sola analisi con il tempo fissato.

2 Analisi trasversale del tasso di turisticità

Analisi del 2000

Come prima cosa è stato creato un dataframe con i valori che le variabili hanno assunto nell'anno 2000, il primo anno di cui si dispongono i dati.
Creato il dataframe, si presenta così:

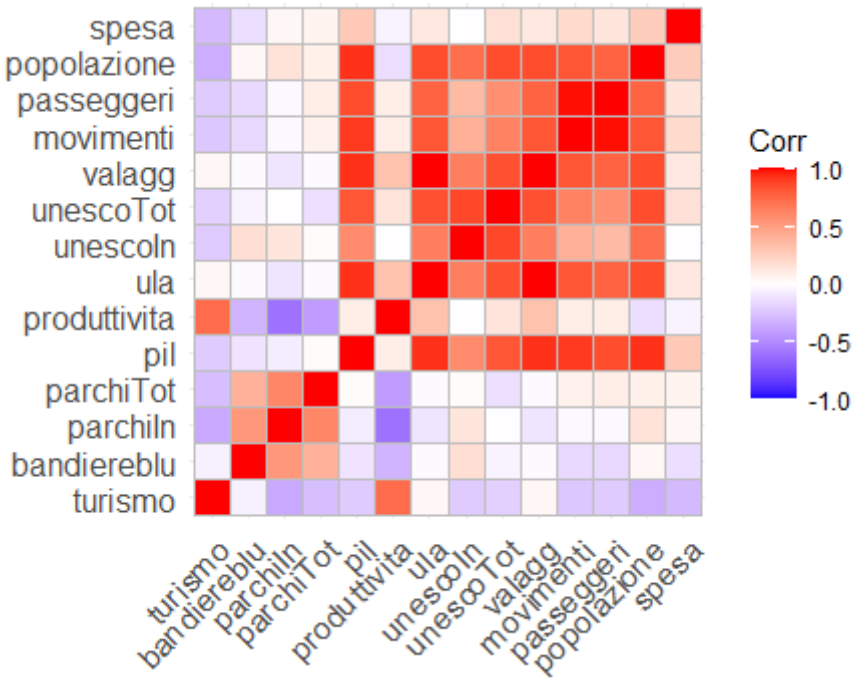
```
head(data2000)
```

	Regioni	turismo	bandiereblu	parchiIn	parchiTot	pil	produttivita
1	Abruzzo	4.978879	14	1	3	24134.8	38.12199
2	Basilicata	2.433604	5	1	2	9717.5	37.13942
3	Calabria	3.104766	17	2	3	26040.3	36.70809
4	Campania	3.623347	18	2	2	82803.6	32.50064
5	EmiliaRomagna	9.202175	9	0	2	106437.8	45.30788
6	FriuliVeneziaGiulia	7.807604	2	0	0	27613.4	53.79149

	ula	unescoIn	unescoTot	valagg	movimenti	passeggeri	popolazione	spesa
1	768.2106	0	1	768.2106	9940	114024	1261134	172.45994
2	235.1328	1	2	235.1328	0	0	601448	20.56837
3	844.5830	0	1	844.5830	18441	1376383	2028007	813.84245
4	2529.3953	5	6	2529.3953	62494	4136508	5717191	147.98037
5	3522.6333	4	5	3522.6333	87276	3896973	3945406	316.65399
6	1071.5131	1	5	1071.5131	19045	574665	1178281	559.57870

Inizialmente si osserva una versione semplificata del correlogramma, con il solo scopo di avere un'idea del comportamento che assumono le variabili in relazione le une con le altre:

```
ggcorrplot(cor(data2000[,-1]))
```



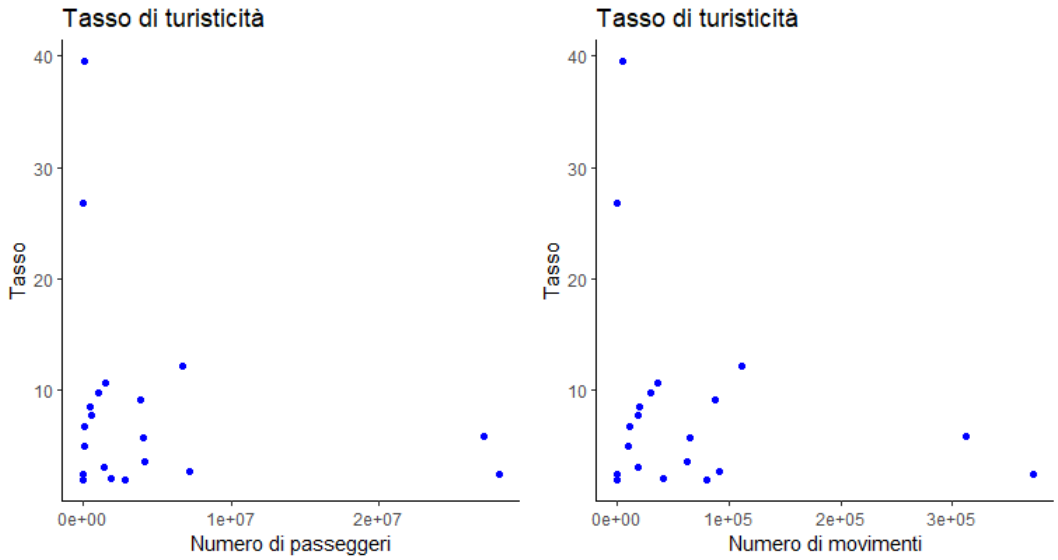
Regressione lineare semplice

Per la regressione lineare semplice sono state candidate due variabili molto simili tra loro: i movimenti aerei e il numero di passeggeri totali. Entrambe le variabili consentono di costruire un modello del tasso di turisticità che può essere interpretato mediante il traffico aereo di una regione.
Il modello che si vuole esprimere in questa sezione descrive un fenomeno del tipo:

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i \text{ con } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In cui Y_i è la variabile risposta, ovvero il tasso di turisticità (o una sua trasformazione), e x_i è la variabile esplicativa, ovvero il numero di movimenti o di passeggeri (o una loro trasformazione). I coefficienti β_1 e β_2 sono invece ignoti, e l'obiettivo del modello sarà stimare questi ultimi; ϵ invece è la componente erratica.

Come prima cosa si osservano i due diagrammi di dispersione:



Graficamente è evidente quanto le due variabili siano simili e interscambiabili per la costruzione di un modello. Per una scelta puramente empirica, dettata anche dalle modifiche effettuate in seguito, si decide di costruire il modello sul numero di movimenti.

Il primo modello è il modello di regressione semplice, quindi in primo luogo si osserva cosa succede effettuando una regressione senza trasformare le variabili[1]:

```
movModel <- lm(turismo ~ movimenti, data = data2000)
```

Il modello ottenuto è il seguente:

$$\hat{y}_i = 9.914 - 2.186 * 10^{-5} x_i$$

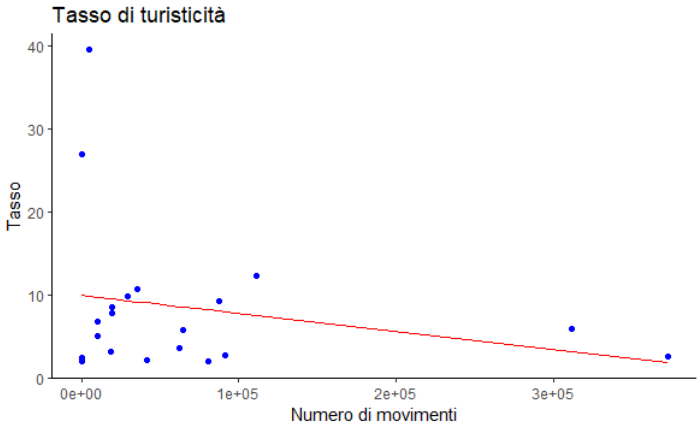
e, già come modello in sé, si osserva come non sia un buon modello. Infatti, l'ordine di grandezza del coefficiente $\hat{\beta}_2$ è molto basso nonostante i dati sui movimenti siano dell'ordine di 10^5.
Il summary del modello ci fornisce ulteriori informazioni che confermano la non bontà del modello: in primo luogo l'R^2 del modello, ovvero il rapporto tra la devianza spiegata dal modello e la devianza totale, è molto basso:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 0.05554$$

Inoltre, un'altra informazione ricavabile sal summary che va a sfavore del modello, e che conferma quanto detto rispetto al coefficiente β_2, è il p-value del test di significatività di β_2 che vale 0.31719.

[1] Summary completo in tabella 1

Come ultima conferma si può visualizzare graficamente il modello:



Il grafico è la conferma definitiva che il modello non funziona, ma allo stesso tempo è un buon punto di partenza per capire quali trasformazioni possono essere efficaci.

Dopo averle provate tutte, la soluzione cui si arriva è che il modello migliore è il modello in cui la variabile risposta non subisce alcuna trasformazione, mentre la variabile esplicativa viene trasformata come reciproco della radice:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \frac{1}{\sqrt{x_i}}$$

Questo modello porta con sé un problema: i dati di partenza sono pochi (solo 20), ed applicando questa trasformazione si vanno a perdere 3 regioni su 20 in quanto non dispongono di un aeroporto.
Tuttavia questo modello è nettamente migliore alle altre trasformazioni, quindi solamente per il modello lineare semplice verrà applicata questa trasformazione.

Si applica quindi il modello in R[2]:

```
movModel2 <- lm(turismo ~ 1/sqrt(movimenti), data = data2000)
```

Il modello ottenuto è il seguente:

$$\hat{y}_i = -1.518 + 1695.628 x_i$$

Anche in questo caso il modello non è pienamente soddisfacente; nel summary si trovano informazioni utili in tal senso: lo standard error di β_1 è 3.440, più del doppio del valore del coefficiente; infatti, il p-value di questo coefficiente è di 0.66540. Il coefficiente β_2, invece, risulta significativo, con uno standard error di 531.044 (che, per quanto grande, è "piccolo" rispetto al valore che assume il coefficiente) e un p-value di 0.00605.
A questo punto conviene costruire un modello in cui non è inclusa l'intercetta[3]:

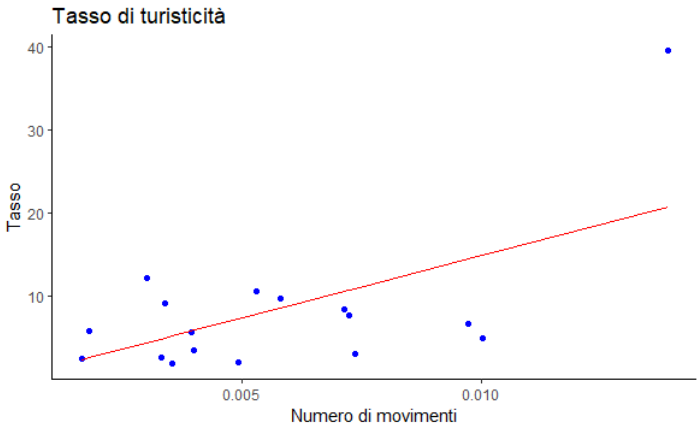
```
movModel4 <- lm(turismo ~ 1/sqrt(movimenti) -1, data = data2000)
```

Ottenendo così un modello del tipo:

$$\hat{y}_i = 1491.4 x_i$$

Questo modello è privo di intercetta, tuttavia l'R^2 è nettamente maggiore del precedente: il modello precedente aveva un R^2 di 0.4047, mentre questo di 0.6839.

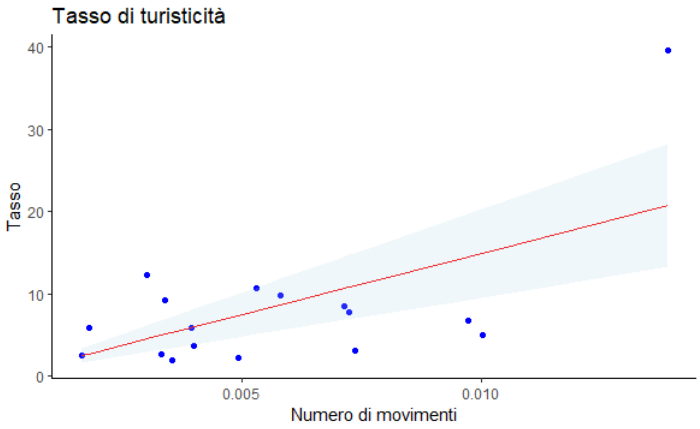
[2] Il summary completo in tabella 2
[3] Summary in tabella 3



Graficamente, questo modello sembra essere accettabile.
Una breve occhiata agli intervalli di confidenza:

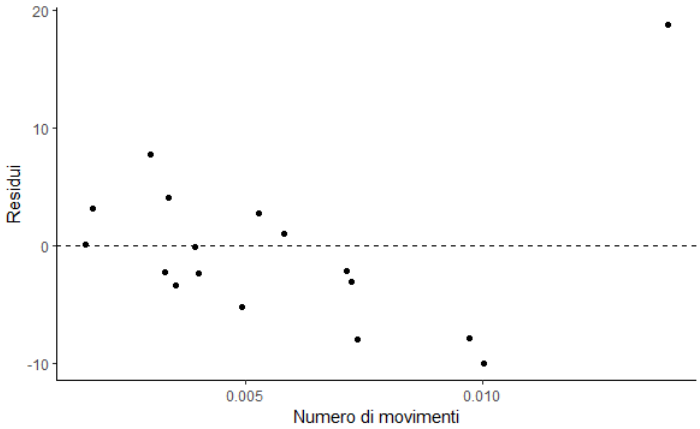
```
confint(movModel4)
```

```
                2.5 %    97.5 %
movimenti  954.0638  2028.694
```

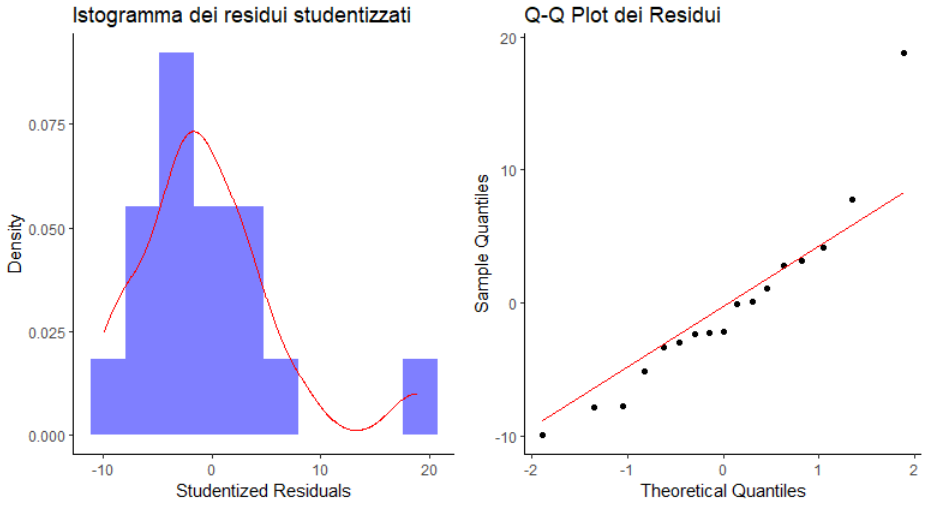


Analisi dei residui

I residui del modello si presentano così:



Si nota subito come non siano dei residui buoni, si vede chiaramente un andamento funzionale. Altrettanto evidente è che il problema è dato dal punto anomalo (Trentino Alto Adige), una regione con un tasso di turisticità molto elevato (quasi 40) e un numero molto basso (ma non nullo) di movimenti aerei. Conferma di ciò si trova anche nei due seguenti grafici:



In questi grafici diventa ancor più evidente come i residui non siano accettabili a causa del valore estremo del Trentino Alto Adige.
Un test analitico che si può utilizzare per verificare la normalità dei residui è il test di Shapiro-Wilk: è un test che viene utilizzato per verificare se il campione di dati proviene da una distribuzione normale; l'ipotesi nulla è che i dati seguano una distribuzione normale.

```
shapiro.test(resid(movModel4))
```

```
        Shapiro-Wilk normality test

data:  resid(movModel4)
W = 0.90577, p-value = 0.08488
```

Il test non rifiuta l'ipotesi di normalità al livello di confidenza del 5%. Tuttavia il risultato è coerente con i grafici, poiché ad eccezione del punto estremo i residui sembrano avere una distribuzione che può essere considerata normale.

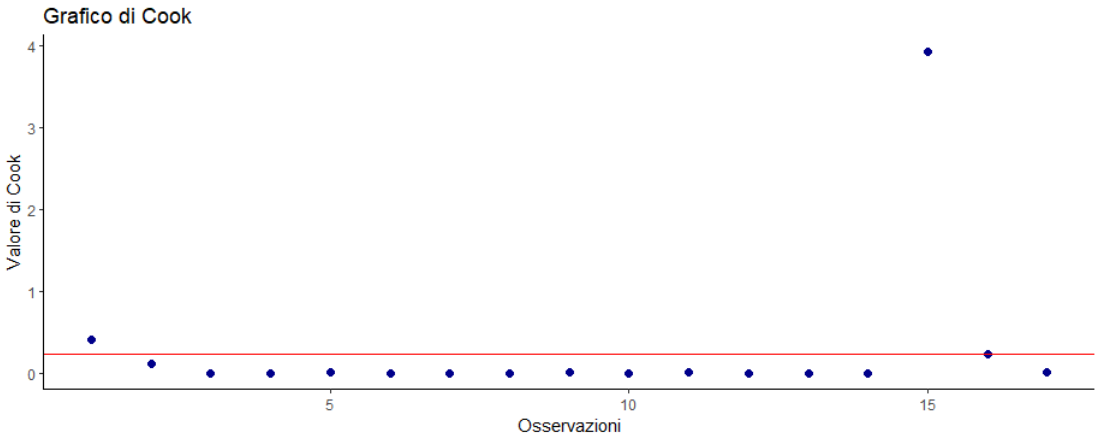
Il modello senza il Trentino

Soprattutto con l'analisi dei residui, è diventato evidente come il Trentino Alto Adige sia un dato non coerente con il resto del dataset; in particolare, è un punto di leva, ovvero un punto che ha un effetto significativo sull'andamento della regressione.
Per individuare i punti di leva si può ricorrere al grafico di Cook: questo mostra il cambiamento nel valore dei coefficienti di regressione quando ogni osservazione viene eliminata dalla regressione.
Nel grafico seguente sono rappresentate le distanze di Cook, calcolate così:

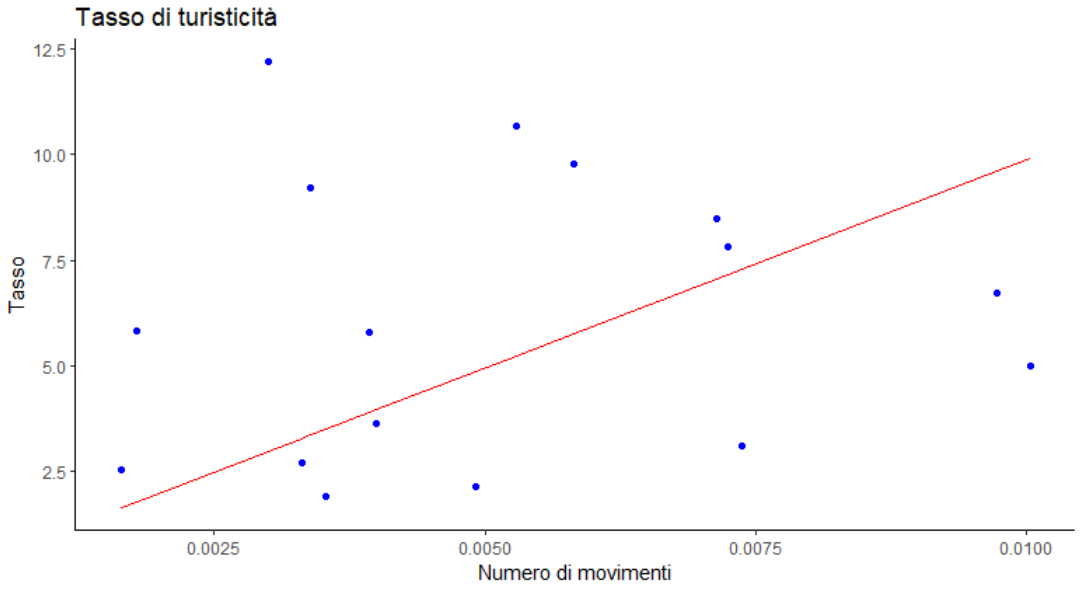
$$D_i = \frac{1}{ps^2}(\hat{\beta} - \hat{\beta}_{-i})^T X^T X (\hat{\beta} - \hat{\beta}_{-i})$$

Sul grafico che segue vengono stampate le distanze di Cook di ciascun dato e una soglia fissata a

$$\frac{4}{n-p} = \frac{4}{17-1} = \frac{1}{4}$$

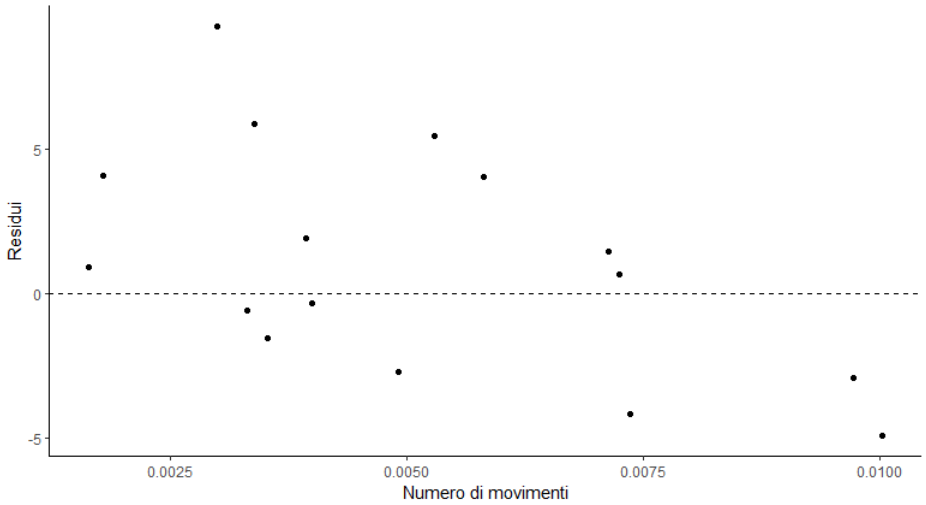


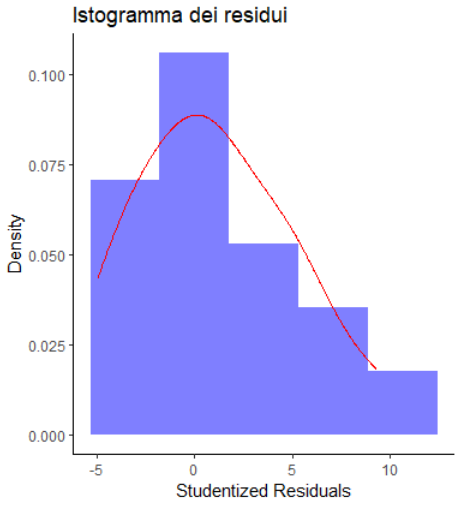
Appare evidente che il punto è di leva (normalmente si possono considerare valori critici i valori > 0.5 o > 1). Pertanto, ora, anche se verrà perso un ulteriore dato, viene rimosso il Trentino Alto Adige dal dataset[4].

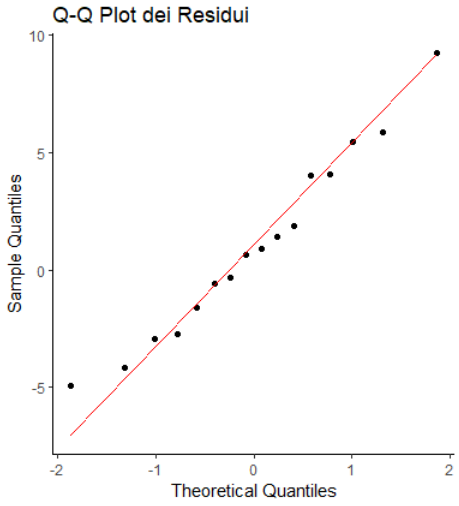


Si può osservare come, con la rimozione del dato anomalo, la scala del grafico sia notevolmente cambiata e i punti sembrino una nube casuale.
Un primo confronto che si può effettuare con il modello precedente è l'R^2: togliendo il Trentino Alto Adige, questo è rimasto praticamente invariato. Tuttavia, una misura che è cambiata molto, è il coefficiente β_2: è passato da 1491.4 a 989.1, mettendo in evidenza l'impatto che aveva il dato del Trentino sull'intero modello.
Seguono i residui:

[4]Summary a tabella 4







Ed infine il test di Shapiro-Wilk:

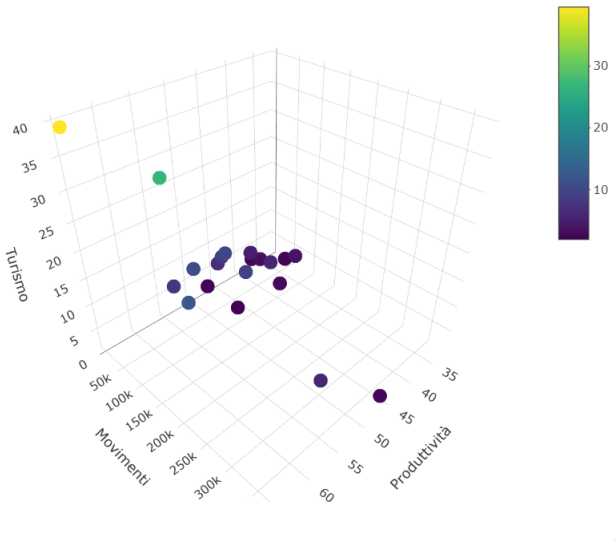
```
shapiro.test(resid(movModel4))
```

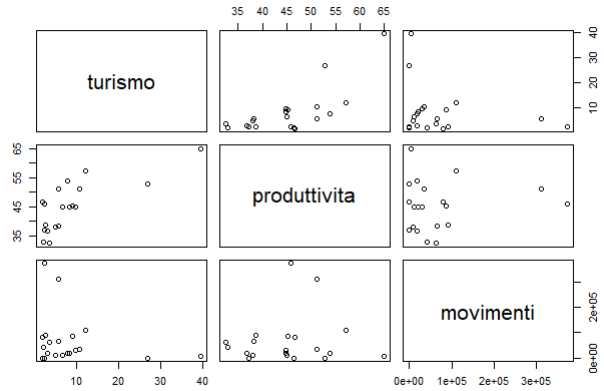
```
        Shapiro-Wilk normality test

data:  resid(movModel4)
W = 0.97371, p-value = 0.8946
```

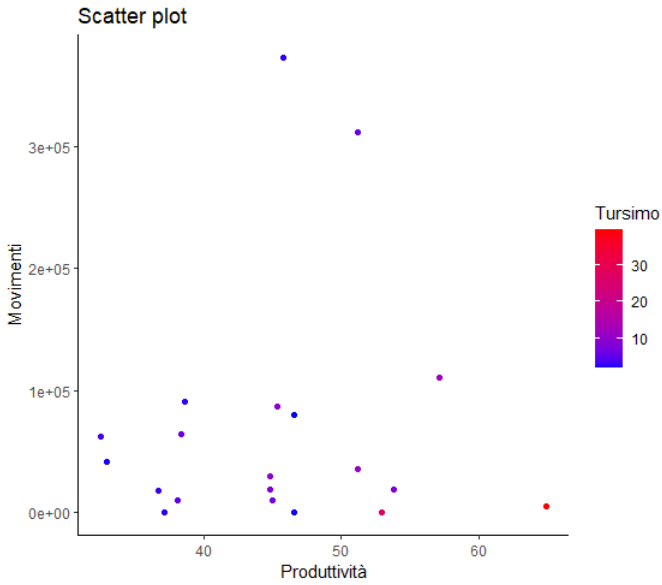
Osservando i residui si può concludere che la rimozione del trentino dal modello ha effettivamente portato ad un miglioramento, sicuramente ora si può accettare più facilmente l'ipotesi di normalità dei residui (sebbene l'istogramma, ma generalmente il confronto della densità con l'istogramma funziona meglio con campioni più ampi).

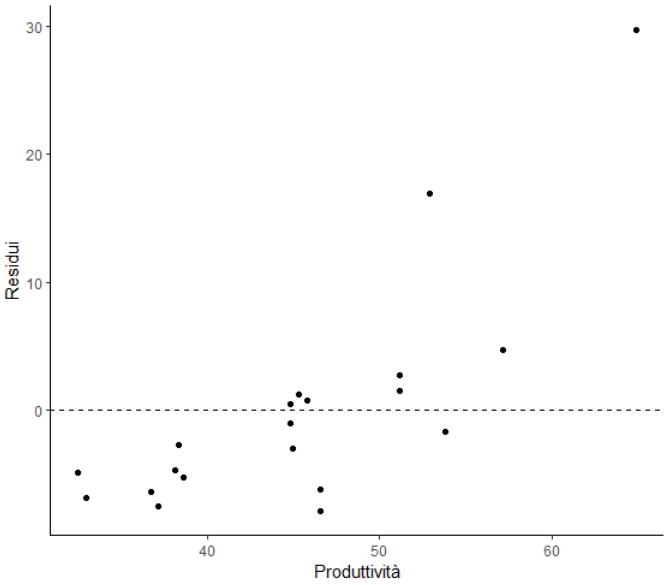
Regressione multipla

Per creare un modello di regressione multipla, come prima cosa è necessario aggiungere una variabile al modello. Si nota dal correlogramma stampato in precedenza che la produttività del lavoro nel settore turistico ha una forte correlazione positiva, quindi potrebbe essere una buona variabile da aggiungere al modello. Come prima cosa è interessante vedere lo scatterplot 3d del tasso di turismo con i movimenti aerei e la produttività del lavoro:









I grafici sembrano buoni, in particolare l'ultimo che rappresenta i residui del modello di regressione lineare semplice del tasso di turisticità sul numero di movimenti, stampati sull'asse della produttività. Si può infatti notare un evidente andamento funzionale.
A questo punto ci sono i presupposti per la creazione di un modello:

```
promovModel <- lm(turismo ~ produttivita + movimenti, data = data2000)
```

Il modello generato è il seguente[5]:

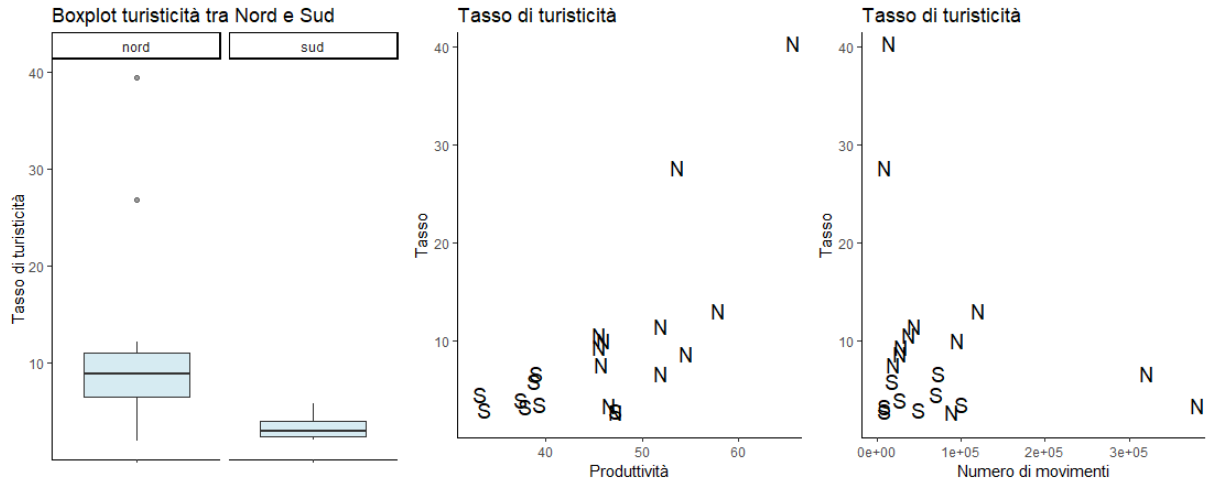
$$\hat{y}_i = -27.88 + 0.8453 x_{i1} - 2.806 * 10^{-5} x_{i2}$$

Si nota subito come l'R^2 corretto sia diminuito rispetto al modello studiato per la regressione lineare semplice (passa da 0.6491 a 0.5875), tuttavia è naturale sia così poiché il nuovo modello comprende 4 dati che sono stati scartati nel modello precedente (Trentino Alto Adige, Valle d'Aosta, Molise e Basilicata) e questi 4 valori sono rappresentativi delle piccole regioni italiane, che nel modello precedente non erano opportunamente rappresentate.

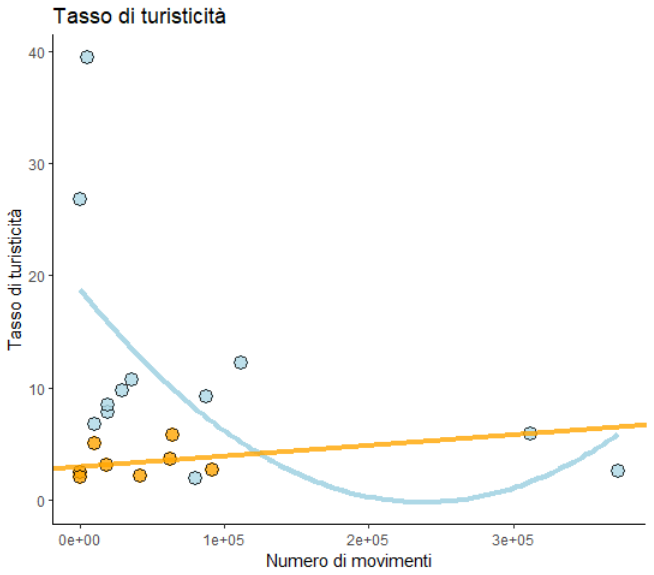
[5]Summary in tabella 5

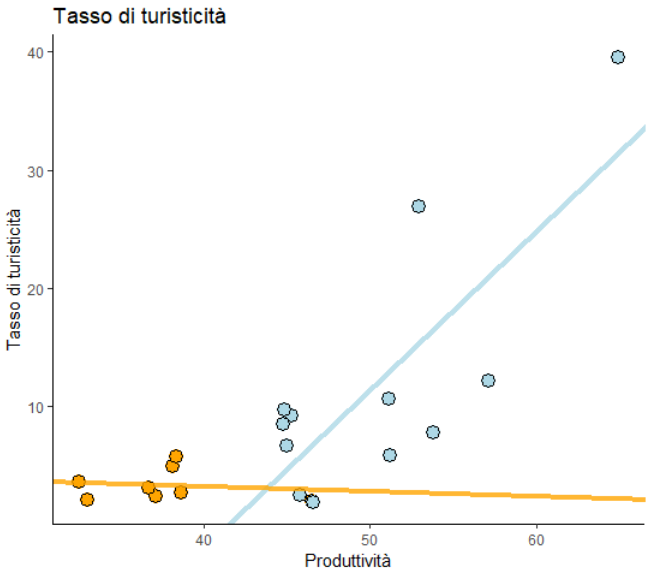
Il coefficiente della produttività è estremamente significativo, mentre quello dei movimenti ha un p-value di 0.05666. Tuttavia, valutando anche il modello con la sola produttività, si decide di mantenere comunque i movimenti aerei nel modello.

Interessante è osservare come i modelli con il valore aggiunto sul lavoro nel settore del turismo e le unità di lavoro nel settore del turismo, rispettivamente numeratore e denominatore della produttività, risultino poco descrittivi del fenomeno, sia graficamente, sia per quanto riguarda l'R^2 e la significatività delle variabili.
Dopo aver provato manualmente tutte le variabili e aver appurato che risultano tutte non significative rispetto al modello con produttività e movimenti, tale modello diviene il modello migliore.
Come ultimo tentativo si prova a dividere le regioni in Nord-Italia e Sud-Italia, creando una variabile qualitativa. L'idea che si cela dietro la creazione di questa variabile è che il divario dei prezzi è molto elevato tra nord e sud, e variabili come la spesa potrebbero non essere risultate significative o aver descritto il modello in modo sbagliato a causa di questo fenomeno.
La suddivisione adottata è in mezzogiorno (Sud Italia e isole) e centro-nord (Centro, Nordovest e Nordest Italia), in quanto utilizzata in un report della Banca d'Italia in cui vengono affrontati i divari economici tra il nord e il sud della penisola.



I grafici mostrano perfettamente l'evidente scarto tra le due grandi ripartizioni d'Italia.
Inoltre è sufficiente effettuare due regressioni, una per i dati del nord Italia e una per i dati del sud, per rendersi conto che le variabili esplicative sono estremamente condizionate dalla posizione geografica, sia nell'intercetta che nel coefficiente.
Nel grafico con il numero di movimenti è stata effettuata una regressione di secondo grado, non funzionando bene la regressione di primo grado.





A questo punto si può formulare un nuovo modello contenente le interazioni con la variabile dicotomica NS (che assume valore "nord" se la regione appartiene al nord Italia, "sud" se appartiene al sud Italia).
In realtà si osserva che nel modello in cui l'interazione nord - sud è presente con entrambe le variabili, l'interazione con i movimenti risulta poco significativa. Il modello finale è quindi[6]:

```
Mymodel <- lm(turismo ~ movimenti + produttivita*NS, data = data2000)
```

Il modello generato è il seguente:

$$\hat{y}_i = -48.81 - 2.559 * 10^{-5} x_{1i} + 1.251 x_{2i} + 57.50 x_{3i} - 1.390 x_{2i} x_{3i}$$

Si possono fare delle considerazioni: l'R^2 corretto è relativamente alto, vale 0.6823 (ed è l'R^2 maggiore tra tutti i modelli provati, anche di quelli cui non è stato mostrato il summary). I coefficienti sono tutti significativi al livello $\alpha = 0.05$ ad eccezione del coefficiente dei movimenti aerei, con un p-value di 0.06; tuttavia, si può osservare che togliendo i movimenti dal modello, il fit peggiora discretamente.
Come ultima cosa, si può visualizzare l'AIC (Akaike Information Criterion):

```
> AIC(Mymodel)
[1] 129.1804
```

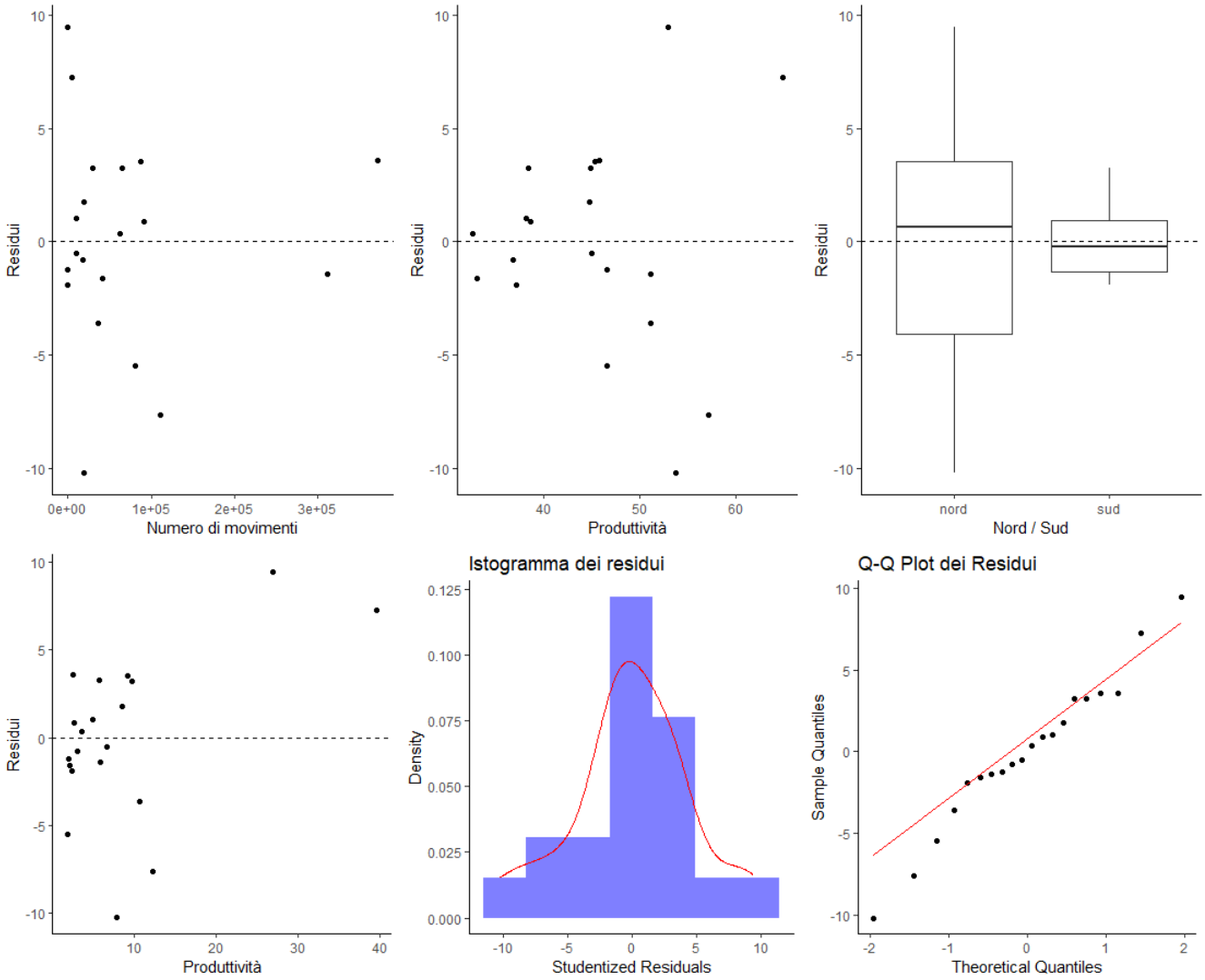
L'AIC tornerà utile più avanti quando si tratterà un modello ottenuto mediante step-AIC.

In seguito al modello ottenuto, sono seguiti ulteriori tentativi di aggiungere o sostituire variabili esplicative non altrettanto soddisfacenti che hanno portato ad accettare l'ultimo modello ottenuto.

[6] Summary in tabella 6

Analisi dei residui

Si visualizzano graficamente i residui rispetto a ciascuna variabile:



Dalla visualizzazione grafica dei residui la linearità sembra accettabile, l'eteroschedasticità e la normalità invece sono pià difficili da accettare.
Si cerca conferma nel test di Shapiro-Wilk:

```
> shapiro.test(resid(Mymodel))

        Shapiro-Wilk normality test

data:  resid(Mymodel)
W = 0.97511, p-value = 0.8568
```

Il test di Shapiro-Wilk porta a non rifiutare l'ipotesi di normalità, tuttavia i grafici lasciano ben più perplessità.

Modello con lo step-AIC

Un'alternativa per la costruzione di un modello è l'utilizzo dello step-AIC:

```
StepModel <- stepAIC(AllInModel, direction = "both")
```

Lo step AIC è un sistema che confronta diversi modelli candidati selezionando il migliore secondo il criterio dell'AIC. L'Akaike Information Criterion si calcola nel seguente modo:

$$AIC = -2\hat{\ell}_{\mathcal{M}_k} + 2k$$

Con $\hat{\ell}_{\mathcal{M}_k}$ funzione di verosimiglianza massimizzata per il modello $\mathcal{M}_k$ con k termini.
L'AIC dà una misura della bontà di adattamento di un modello, ovvero quanto il modello è in grado di spiegare il comportamento della variabile risposta.
Il modello ottenuto dallo step-AIC è:

$$\hat{y}_i = -45.35 + 1.247x_{1i} - 2.968x_{2i} - 7.182 * 10^{-5}x_{3i} + 5.483 * 10^{-6}x_{4i} - 8.474 * 10^{-3}x_{5i}$$

Con x_1 produttività, x_2 numero di siti Unesco nella regione, x_3 numero di movimenti aerei, x_4 popolazione residente nella regione, x_5 spesa totale effetutata dai turisti nella regione.

Il modello ottenuto via step-AIC è particolarmente interessante[7]: l'AIC del modello è estremamente più basso del modello selezionato in precedenza (65.89 contro 129.18) e pure l'R^2 corretto risulta essre più alto (0.7545 contro 0.6823); e nonostante ciò, anche le variabili sono più significative di quelle del modello ottenuto in precedenza.
Tuttavia, il modello ottenuto via step-AIC è difficilmente interpretabile e, nonostante sia il migliore per quanto riguarda la bontà di adattamento, 5 variabili potrebbero risultare troppe.
Nasce la necessità di misurare il principio di parsimonia, valutata in modo più preciso dal BIC. Il Bayesian Information Criterion (BIC) si calcola nel seguente modo:

$$BIC = -2\hat{\ell}_{\mathcal{M}_k} + k\ln(n)$$

Si calcola quindi il BIC dei due modelli:

```
> BIC(Mymodel)
[1] 135.1548
> BIC(StepModel)
[1] 131.6197
```

Dal confronto dei BIC emerge che i due modelli sono quasi "equivalenti" secondo questo criterio.

Interpretazione dei coefficienti

Il modello di cui si vanno a vedere i coefficienti è il modello individuato con il metodo Foreward:

- $\hat{\beta}_1$: l'interpretazione dell'intercetta è la più forzata, in quanto significa che quando tutte le variabili esplicative assumono valore 0, il tasso di turisticità ha un valore stimato di -48.81.
  Tuttavia questa interpretazione non ha un significato pratico, in quanto la produttività non assume mai valore 0, quindi l'intercetta rappresenta una situazione che non si verifica mai. Infatti il tasso di turisticità è sempre positivo, e non può pertanto assumere il valore -48.81.
- $\hat{\beta}_2$: è il coefficiente della variabile "movimenti", ovvero il numero di aerei che transitano in una regione nell'arco di un anno.
  Il coefficiente è negativo, e può essere interpretato nel seguente modo: le regioni con più movimenti aerei hanno un tasso di turisticità più basso, poiché tendenzialmente le regioni con molti abitanti hanno più traffico aereo, e la popolazione è il denominatore del tasso di turisticità. Probabilmente il traffico aereo implica un aumento nella popolazione maggiore rispetto all'aumento del numero di turisti.
- $\hat{\beta}_3$: è il coefficiente della variabile "produttivita": indica la produttività del lavoro nel settore turistico.
  Il coefficiente ha valore positivo, ciò vuol dire che l'aumento della produttività del lavoro nel settore del turismo implica un aumento del tasso di turisticità. In particolare questa relazione vale per le regioni del Nord-Italia.

[7] Summary in tabella 7

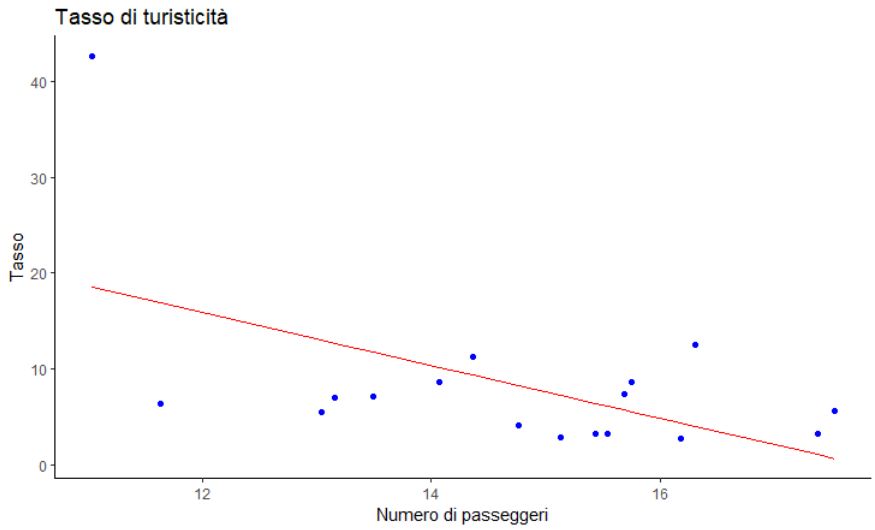
- $\hat{\beta}_4$: è il coefficiente della variabile "NS", che assume valore 0 se la regione appartiene al Nord-Italia e 1 se appartiene al Sud-Italia.
  Il coefficiente è positivo, e rappresenta l'aumento del tasso di turisticità che si osserva passando da una regione del Nord-Italia ad una del Sud-Italia.
- $\hat{\beta}_5$: è il coefficiente del prodotto tra le variabili "produttivita" e "NS".
  Il coefficiente rappresenta la variazione negativa del coefficiente della produttività che si osserva per le regioni del Sud-Italia.
  Pertanto, per una regione del Nord, all'aumento della produttività aumenta anche il tasso di turisticità; per le regioni del Sud, invece, essendo $|\hat{\beta}_5| > |\hat{\beta}_3|$, all'aumento della produttività il tasso di turisticità diminuisce.

E negli altri anni?

Sono stati calcolati i modelli per altri 3 anni con gli stessi metodi: 2010, 2019 e 2020.
Seguono solamente i risultati.

- Regressione lineare semplice nel 2010:
  La variabile esplicativa che meglio spiega il tasso di turisticità nel 2010 è il logaritmo del numero di passeggeri:

$$\text{Turisticità} = 49.045 - 2.763 * \log(\text{Passeggeri})$$

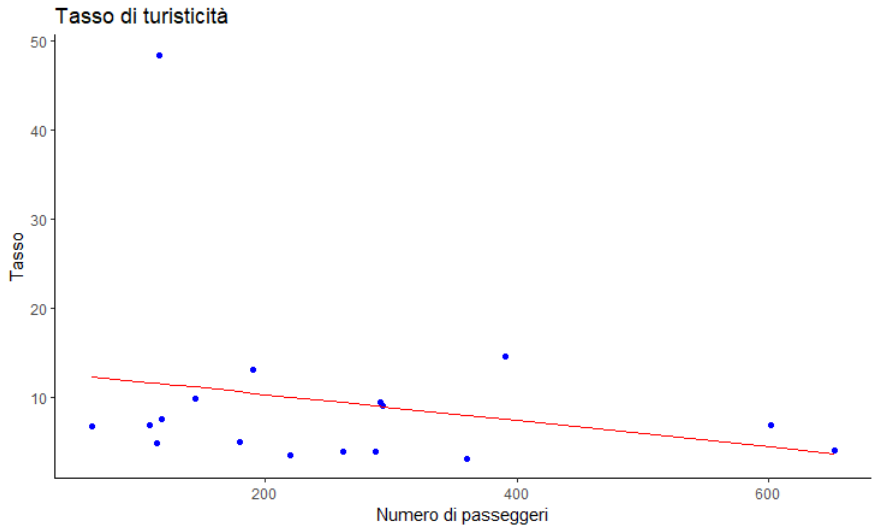


- Regressione lineare multipla nel 2010:
  Le variabili esplicative che spiegano meglio il tasso di turisticità sono PIL e ULA (unità di lavoro nel settore del turismo) senza intercetta, mentre i dati sul traffico aereo sono non significativi in ogni modello:

$$\text{Turisticità} = -3.183e-04 * \text{Pil} + 1.263e-02 * \text{Ula}$$

- Regressione lineare semplice nel 2019:
  La variabile esplicativa nel 2019 torna ad essere la variabile movimenti, ma stavolta nella trasformazione in radice:

$$\text{Turisticità} = 13.20125 - 0.01464 * \sqrt{\text{Movimenti}}$$

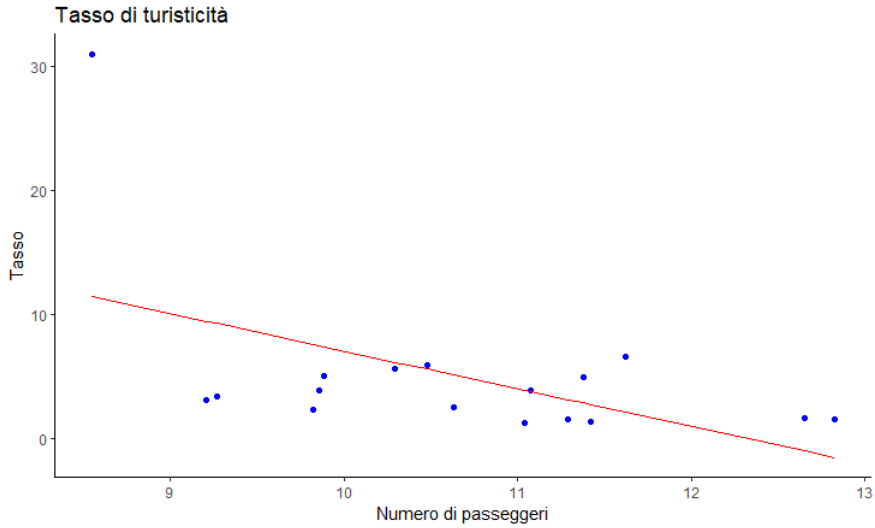


- Regressione lineare multipla nel 2019:
  Le variabili esplicative che spiegano meglio il tasso di turisticità nel 2019 sono il PIL e la variabile indicatrice Nord-Sud:

$$\text{Turisticità} = 1.624e+01 - 4.432e-05 * \text{Pil} - 1.028e+01 * \text{NS}$$

- Regressione semplice nel 2020: La variabile che meglio spiega il tasso di turisticità nel 2020 è il logaritmo del numero di movimenti:

$$\text{Turisticità} = 37.367 - 3.028 * \log(\text{Movimenti})$$



- Regressione lineare multipla nel 2020:
  Le variabili che spiegano il tasso di variabilità nel 2020 sono molto particolari: sono il numero di passeggeri, il numero di movimenti e la variabile indicatrice Nord - Sud:

$$\text{Turisticità} = 1.140e+01 - 1.219e-04 * \text{Movimenti} + 2.434e-06 * \text{Passeggeri} - 9.738e+00 * \text{NS}$$

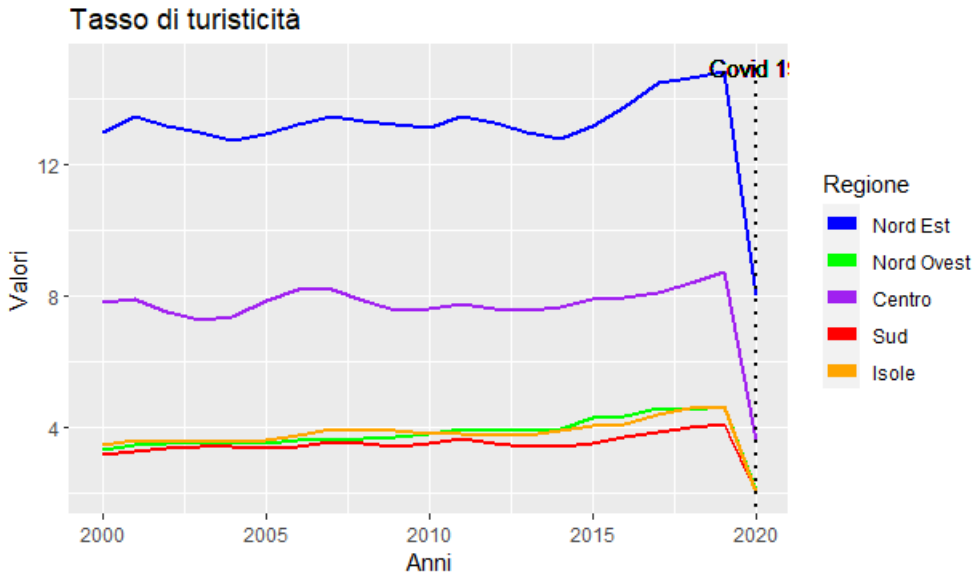
Conclusioni

Dai modelli ricavati nelle precedenti sezioni si possono trarre diverse conclusioni:

- **2000**:
  Nell'analisi sul 2000, alla fine, sono stati individuati 2 modelli quasi equivalenti per il BIC; tuttavia, i due modelli ricavati sono estremamente diversi: il primo, che esprime il tasso di turisticità in funzione di produttività, movimenti aerei e collocazione Nord-Sud di una regione, è molto più interpretabile e consente di giustificare i coefficienti in modo coerente con la realtà (ciò non implica che l'interpretazione sia corretta o che non possano esserci correlazioni spurie); il secondo, che esprime il tasso di turisticità in funzione di produttività, numero di siti Unesco, numero di movimenti aerei, popolazione residente e spesa totale, non sembra di un'interpretazione altrettanto semplice: la produttività e la popolazione residente sono le uniche variabili con coefficiente positivo (e la popolazione è il denominatore del tasso, quindi ha una correlazione negativa matematica con il tasso!), mentre tutte le altre variabili che, intuitivamente, dovrebbero portare un aumento del turismo al loro incremento, invece hanno coefficiente negativo.
- **2010**:
  L'analisi sul modello del 2010 porta a conclusioni molto diverse da quelle sul 2000. Infatti, se i passeggeri riescono a spiegare discretamente il tasso di turisticità in una regressione lineare semplice, nel momento in cui si è passati alla regressione multipla, con l'aggiunta di PIL e ULA, i passeggeri sono diventati non significativi. Il modello è in parte interpretabile: con l'aumento del PIL diminuisce il tasso di turisticità (essendo il PIL non pro capite, un aumento della popolazione comporta un aumento del PIL, ma allo stesso tempo una diminuzione del tasso), mentre all'aumento delle unità lavorative nel settore del turismo aumenta il tasso di turisticità.
- **2019**:
  L'analisi sul modello del 2019 porta ad un risultato nuovamente diverso: la turisticità è spiegata dal PIL, che però stavolta ha coefficiente positivo, interpretabile come "le regioni più ricche attirano più turismo", e dalla variabile indicatrice Nord-Sud, che penalizza le regioni del sud.
  Questo modello è il più particolare tra quelli ottenuti, poiché non entra in gioco nessuna variabile strettamente legata al turismo, ma solo variabili di carattere economico e geografico.
- **2020**:
  Nel modello del 2020 ci si aspettava che le cose cambiassero molto, nonostante disti di solamente un anno dal 2019, studiato anch'esso. Tuttavia l'arrivo del Covid-19 ha completamente rivoluzionato le distribuzioni dei dati.
  Il modello finale è un modello molto particolare, poiché coinvolge contemporaneamente le variabili movimenti e passeggeri: fino al 2020 sono sempre state considerate due variabili con una correlazione troppo elevata per poter essere incluse congiuntamente in un modello, per il legame naturale che intercorre tra il numero di aerei e il numero di passeggeri che ci viaggiano a bordo; tuttavia, nel 2020 sono state attraversate diverse fasi che imponevano delle limitazioni per quanto riguarda la capienza aerea, alternate ad altre fasi (come l'estate) in cui invece la capienza era la massima capienza fisica dell'aereo. Questa continua variazione del rapporto tra numero di voli e di passeggeri ha fatto sì che la correlazione tra le due variabili diminuisca e che siano compatibili per essere inserite congiuntamente in un modello. Addirittura, la variabile movimenti assume segno negativo mentre la variabile passeggeri assume segno positivo.

3 Analisi per serie storiche

In quest'analisi i dati vengono aggregati nelle 5 macro aree dell'Italia. Sono pertanto ridimensionati tutti i dataframe di cui si è parlato prima.
Come prima cosa può essere utile visualizzare graficamente le serie storiche del tasso di turisticità nei 5 gruppi di regioni:



Si osserva immediatamente come il 2020 sia un evento assolutamente estremo per la determinazione di questo tasso. Se il modello non fosse in serie storica, si potrebbe prendere in considerazione l'idea di trascurare il dato del 2020, tuttavia in questo caso la variabile esplicativa temporale ha un ordine e una direzione, pertanto il modello perderebbe completamente la sua capacità previsiva tralasciando l'ultimo dato.

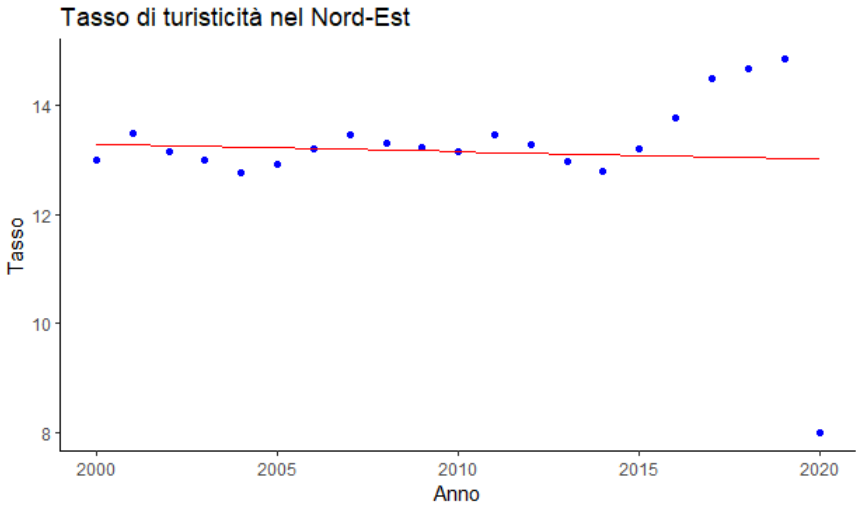
Modello sul Nord-Est

Ai fini della sintesi, la costruzione del modello sul Nord-Est Italia viene seguita in modo approfondito, mentre sulle altre 4 aree, essendo la procedura analoga, ma ne verranno comunque esposti i risultati.
Per quanto riguarda il tasso del Nordest, il primo modello è stato costruito utilizzando come variabile esplicativa la sola variabile temporale:

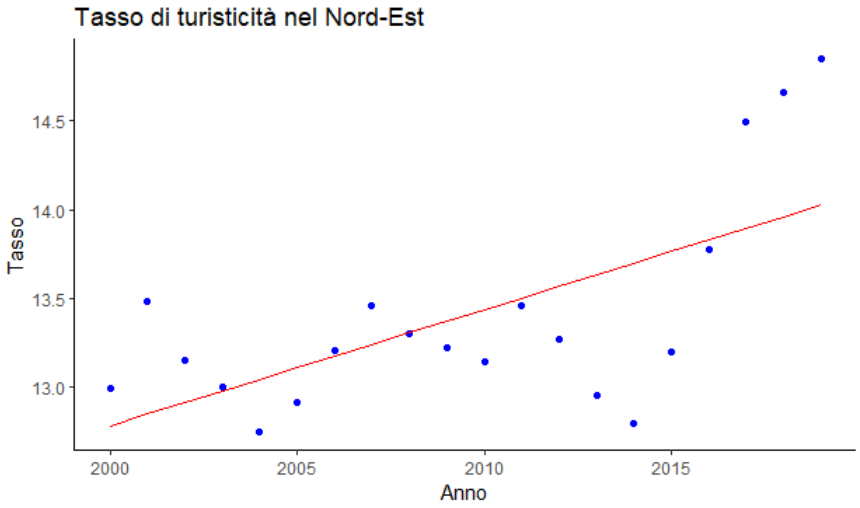
```
TimeModel_NE <- lm(turismo ~ anno, data = NordEst_ts)
```

Si può osservare che nulla funziona in questo modello[8], e il grafico conferma il tutto:

[8] Summary del modello a fine PDF, tabella 8



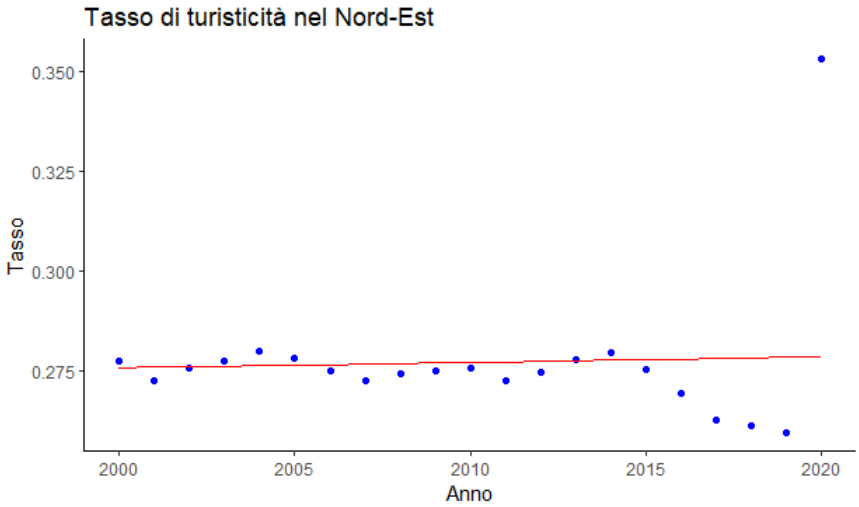
Evidente è che il dato del 2020 sbilancia molto il modello. Né l'intercetta né il coefficiente dell'anno sono significativi, tuttavia rimuovendo l'intercetta il coefficiente lo diventa.
Escludendo il 2020 dai dati, invece, si ottiene un trend lineare di questo tipo[9]:



Appare ancora evidente come il trend lineare non sia il trend migliore per descrivere questo fenomeno, ma sicuramente i risultati sembrano meglio.
Come detto in precedenza, però, studiare il dataframe senza l'ultimo dato permette delle analisi e delle conclusioni molto più limitate.
Si ricerca piuttosto se esiste una trasformazione che permette di arrivare ad un modello migliore.
La trasformazione individuata come migliore è il reciproco della radice della variabile risposta. Quello che risulta è[10]:

```
TimeModel_NE <- lm(1/sqrt(turismo) ~ anno, data = NordEst_ts)
```

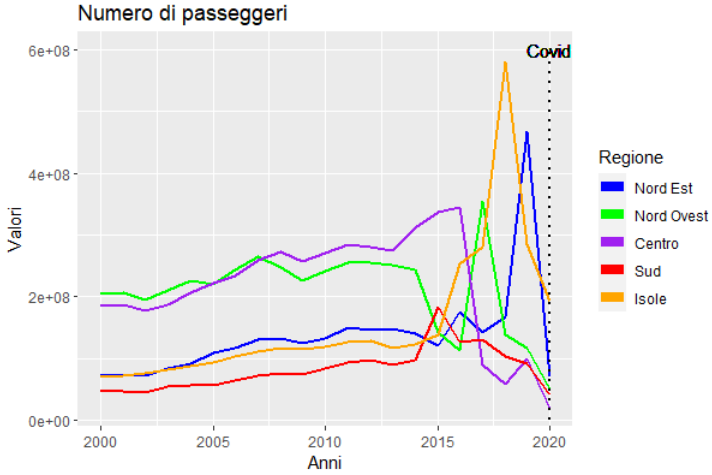
[9] Summary in tabella 9
[10] Summary nella tabella 10

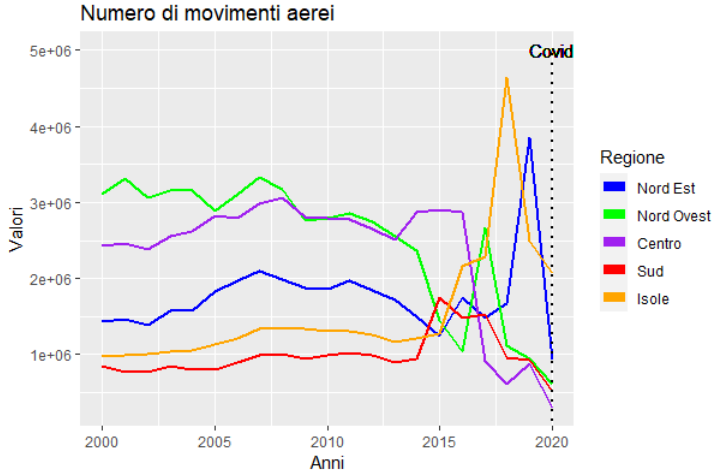


Si può vedere forse un miglioramento nel modello, ma comunque risulta difficile da interpretare e incompleto.

Traffico aereo

Il momento di introdurre una nuova variabile è arrivato: ora inizia la valutazione per capire se includere nel modello la variabile movimenti o la variabile passeggeri.
In primo luogo si osserva il comportamento delle due serie.

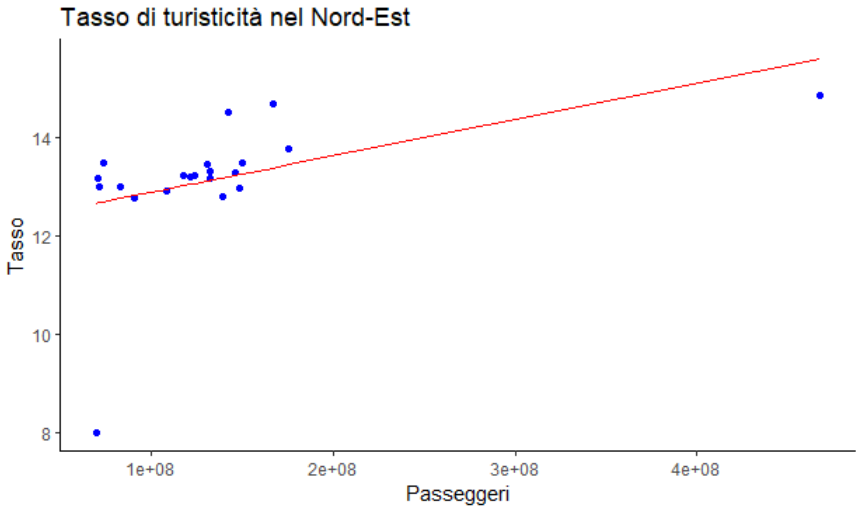




La cosa positiva è che entrambe le variabili hanno avuto un crollo nel 2020, pertanto è più facile che riescano a spiegare il dato del tasso nel 2020.
Si prova a creare entrambi i modelli per confrontarli, prima solo con il tasso, poi anche con il tempo[11].

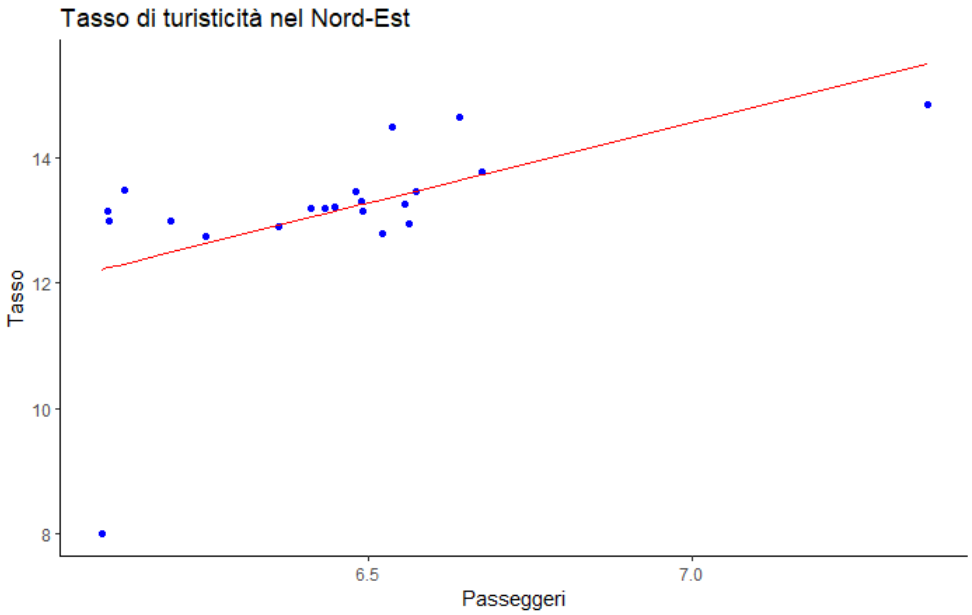
```
PasModel_NE <- lm(turismo ~ passeggeri, data = NordEst_ts)
PasTimeModel_NE <- lm(turismo ~ anno + passeggeri -1, data = NordEst_ts)
MovModel_NE <- lm(turismo ~ movimenti, data = NordEst_ts)
MovTimeModel_NE <- lm(turismo ~ anno + movimenti -1, data = NordEst_ts)
```

[11] Summary nelle tabelle 11,12,13,14

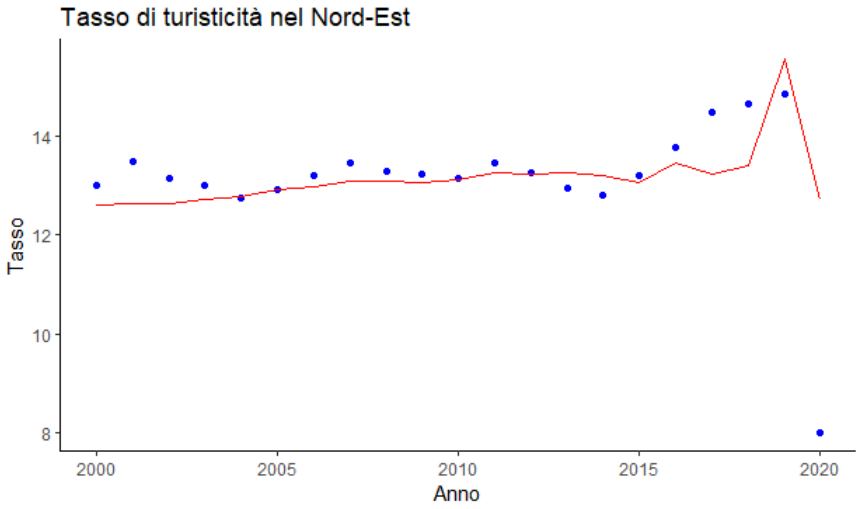


Si nota subito da questo primo grafico come una trasformazione forse potrebbe aiutare.
Trasformando la variabile passeggeri il punto estremo, corrispondente al dato del 2019, dovrebbe ”avvicinarsi” al resto dei punti. La trasformazione migliore è la radice quarta; all'aumentare del grado della radice, la nube di punti inizia ad ”allinearsi” con i due punti estremi, facendoli così divenire meno anomali.

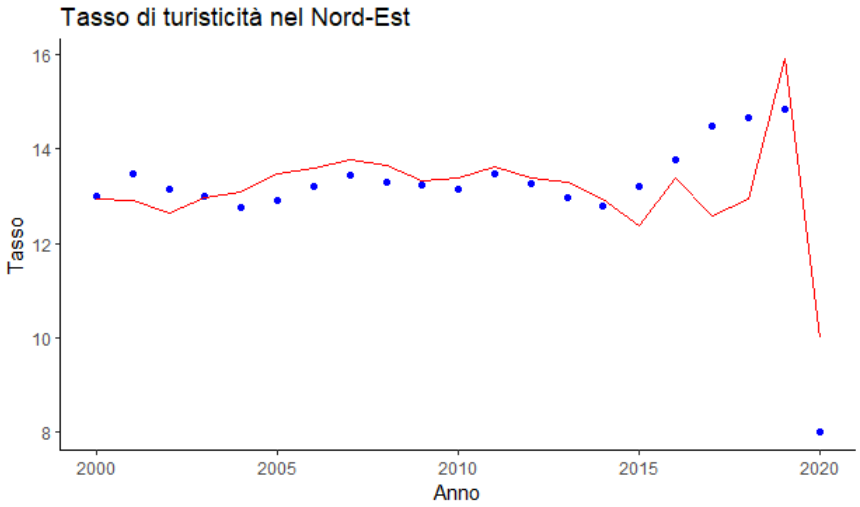
```
PasModel_NE2 <- lm(turismo ~ I(passeggeri^(1/4)) -1, data = NordEst_ts)
```



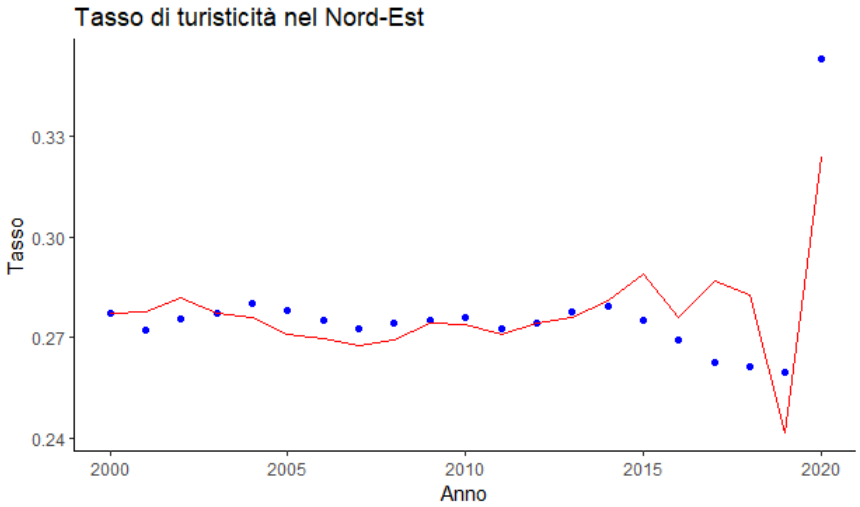
Effettivamente il modello sembra sensibilmente migliore, ma non sembra riuscire a spiegare il crollo del 2020.
A questo punto si visualizza il modello in cui sono inclusi sia il tempo, sia i passeggeri. Per visualizzare graficamente i dati e il modello, si effettua la rappresentazione su due dimensioni dove un asse è rappresentato dalla variabile risposta e l'altro dall'asse temporale.



La variabile passeggeri non è particolarmente significativa, infatti il trend è più o meno lineare se non per l'ultima parte.
Si può provare ad applicare una delle due trasformazioni individuate fin'ora.
Il modello sembra migliorare includendo l'intercetta e trasformando la variabile passeggeri con il logaritmo, anche se l'R^2 diminuisce molto (fino a 0.55); tuttavia i coefficienti sono tutti significativi.

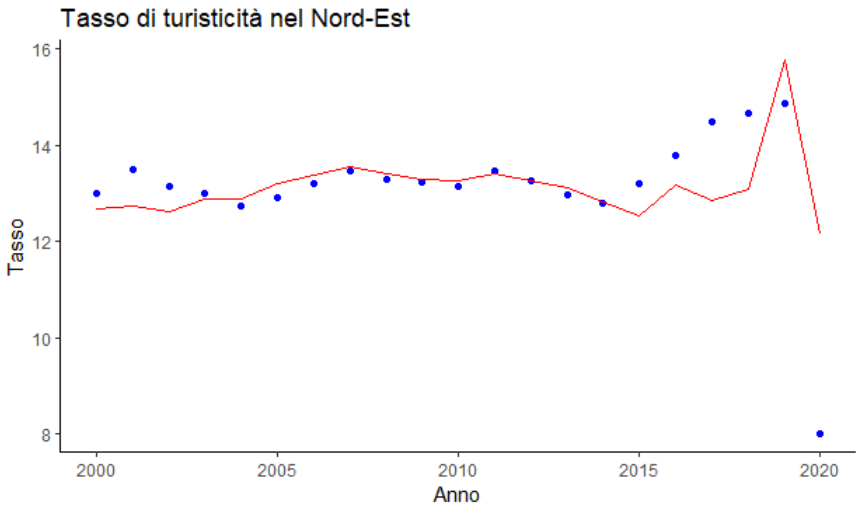


Una trasformazione alternativa, che mantiene i R^2 e significatività dei coefficienti più o meno come la precedente, è il reciproco della radice del tasso di turisticità, mantenendo sempre l'intercetta e il logaritmo dei passeggeri.



Tutti i modelli con i passeggeri hanno difficoltà nella spiegazione dei dati tra il 2016 e il 2018. Gli ultimi due modelli sembrano equivalersi, quindi il modello che verrà mantenuto come migliore con il numero di passeggeri è quello dove la variabile risposta non è trasformata, per facilitare l'interpretazione[12].

A questo punto è tempo di vedere il modello che include i movimenti. Dopo considerazioni analoghe si arriva al modello che include le variabili tempo e movimenti, senza compiere alcuna trasformazione e senza includere l'intercetta.
Il fit finale del modello in questione è il seguente:

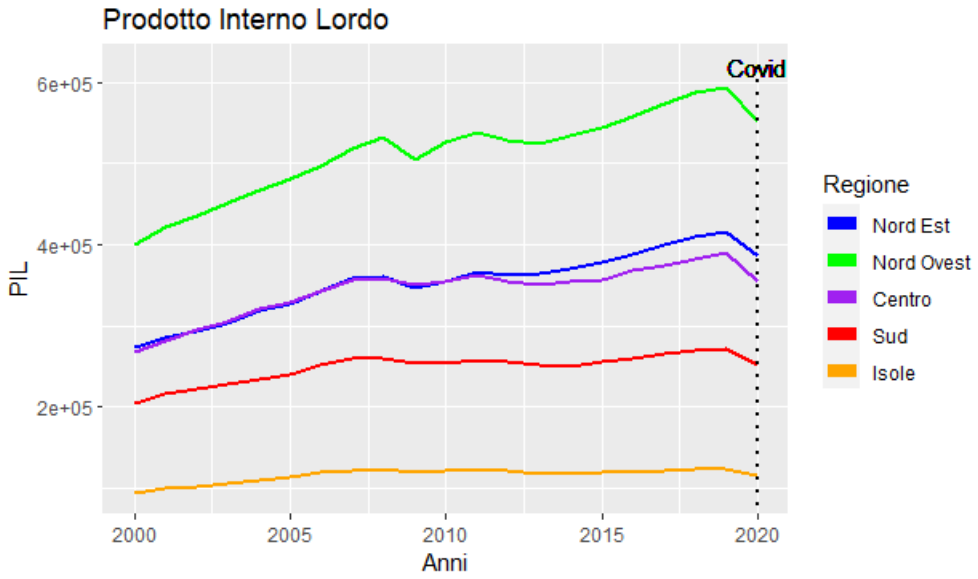


Per ora i modelli che tengono in considerazione il numero di passeggeri e il numero di movimenti si mostrano equivalenti.

[12]Summary in tabella 15

PIL

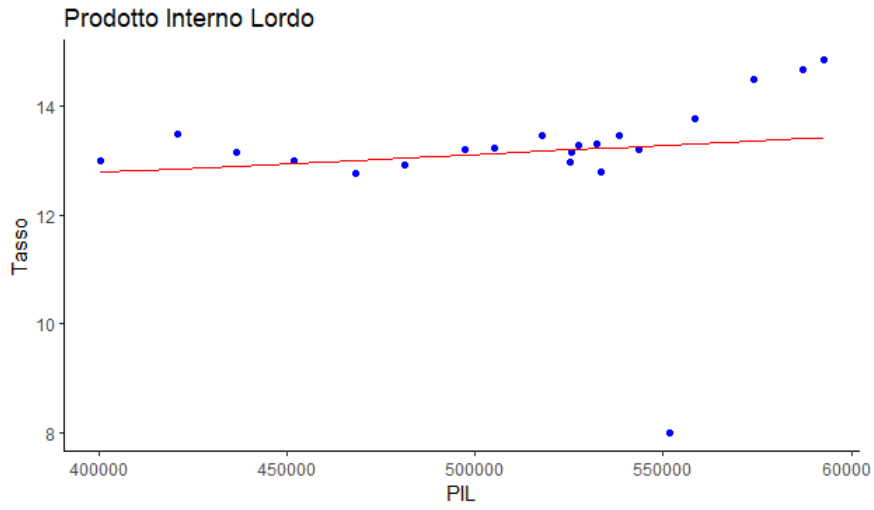
Per arricchire il modello, si prova ad aggiungervi la variabile PIL. Come prima cosa, si osserva il suo comportamento:



Il PIL è una variabile piuttosto lineare, quindi probabilmente, includendo il tempo, il PIL potrebbe risultare poco significativo (e viceversa). Tuttavia il PIL ha un trend di lungo periodo, ad esclusione del 2020 ovviamente, assimilabile a quello del tasso di turisticità. Pertanto, potrebbe avere senso aggiungerlo al modello.

Il primo risultato interessante si ottiene con la solita regressione del tasso di turisticità sul PIL[13]:

```
PilModel_NE <- lm(turismo ~ pil, data = NordEst_ts)
```

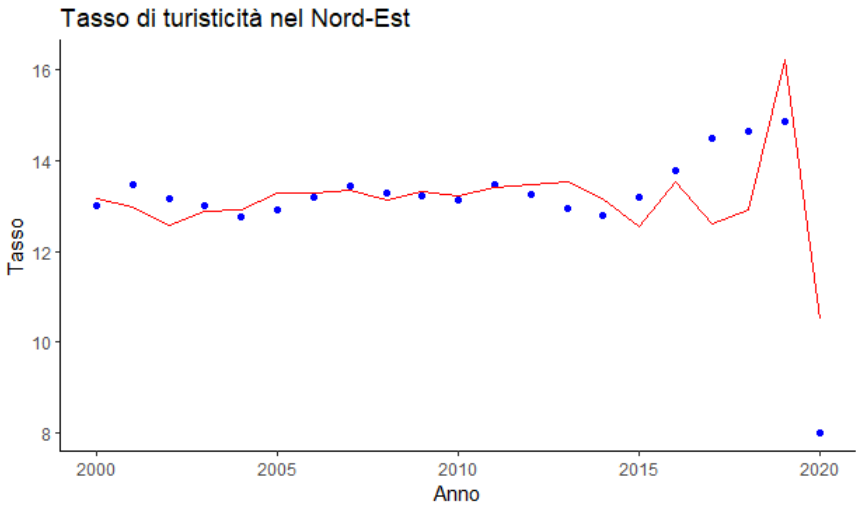


Il problema di questa regressione, apparentemente buona, è che c'è un dato completamente fuori trend, che in un normale database potremmo trattare come tale e considerare un evento anomalo, ma in una serie storica non è possibile fare una semplificazione tale (in particolare poiché il punto è l'ultima misurazione). Il problema risiede nel fatto che il punto anomalo non è un punto di leva, quindi in una prospettiva previsiva il modello con il PIL non risulta essere utile poiché "ignora" il dato sul 2020.

[13]Summary in tabella 16

A questo punto il PIL va aggiunto ai due modelli analizzati in precedenza. Il modello migliore che si ricava aggiungendo il PIL al modello con i passeggeri è il seguente, ovvero quello in cui si aggiunge il PIL senza trasformazioni[14].

```
PilPasTimeModel_NE2 <-lm(turismo ~ anno + log(passeggeri) + pil -1, data = NordEst_ts)
```

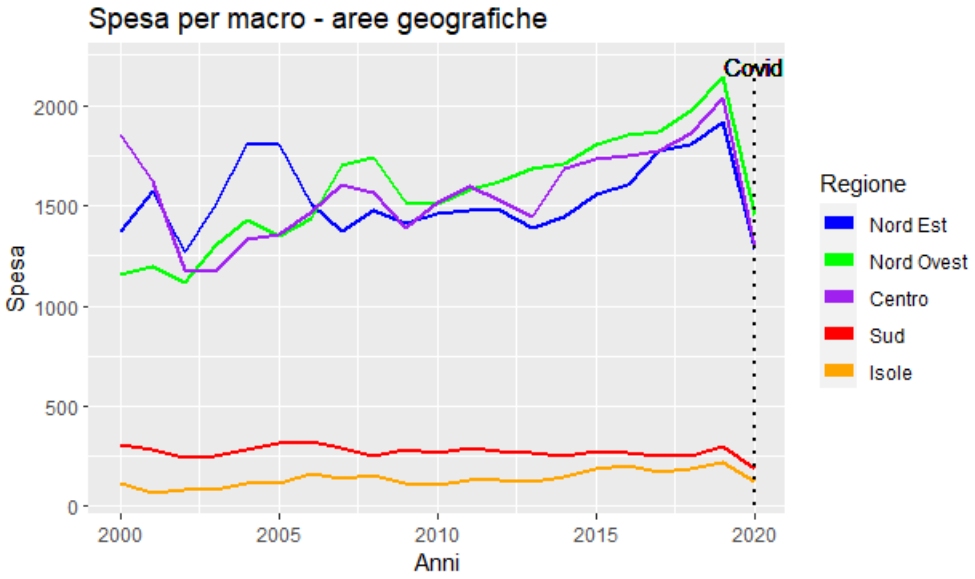


Tuttavia, pur aggiungendo il PIL, il modello non sembra migliorare, quindi ha senso tralasciare la variabile non avendo un effetto significativo.
Per quanto riguarda i movimenti aerei, l'aggiunta del PIL ha dato risultati peggiori, pertanto non viene incluso neanche in quel modello.

Spesa

Il PIL non è stato un buon indicatore per il tasso di turisticità, ma non esclude che la spesa effettuata dai turisti, variabile molto più mirata, non possa esserlo.
In particolare, una caratteristica che dovrebbe favorire la variabile spesa ad entrare tra le esplicative del modello finale è la sua misurazione: infatti i dati sulla spesa dei turisti sono misurati in una serie di categorie di esercizi ricettivi e di ristoro, e sono le stesse categorie esercizi dalle quali vengono estrapolati i dati per la misurazione delle giornate di presenza dei turisti, che è il numeratore della variabile esplicativa.
Come prima cosa si osserva l'andamento della spesa nelle 5 macro aree:

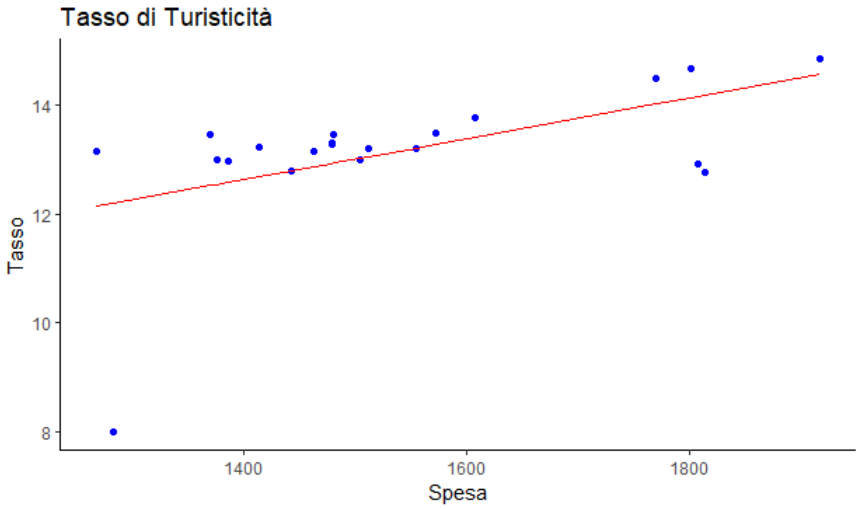
[14] Summary in tabella 17



Questa è la prima variabile ad assumere un comportamento così asimmetrico tra Nord e Sud. Potrebbe quindi anche assumere ruoli diversi nei vari modelli.
La variabile spesa, per le aree del nord Italia (quindi per il Nordest, regione sotto analisi in questo momento), potrebbe spiegare bene il crollo del tasso di turisticità del 2020, in quanto la spesa sembra aver avuto un comportamento simile, almeno graficamente.

Come prima cosa si valuta la regressione del tasso di turisticità sulla spesa[15]:

```
SpeModel_NE <- lm(turismo ~ spesa, data = NordEst_ts)
```



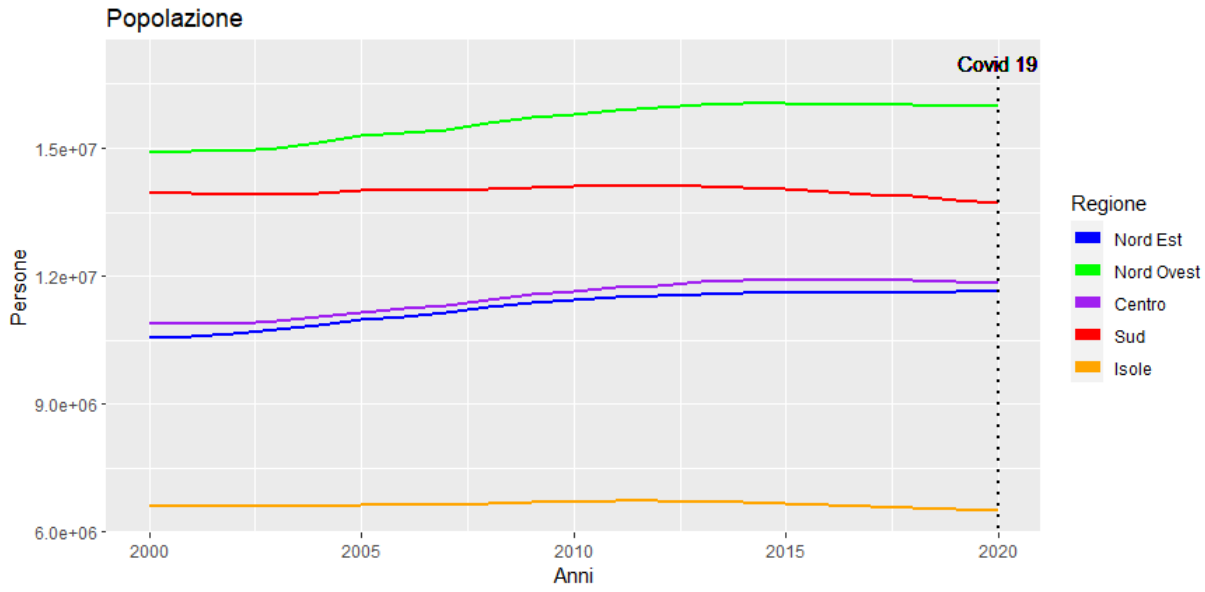
Anche in questo caso, però, analogamente al PIL, è difficile fare sì che la regressione tenga conto del valore estremo del 2020. Tutte le trasformazioni non permettono di ottenere risultati migliori.

La spesa non risulta inoltre incisiva nel modello con i passeggeri, né in quello con i movimenti.
Utilizzata come variabile esplicativa insieme al tempo, senza ricorrere ai dati sul traffico aereo, dà un modello che comunque non è soddisfacente come i precedenti.

[15]Summary in tabella 18

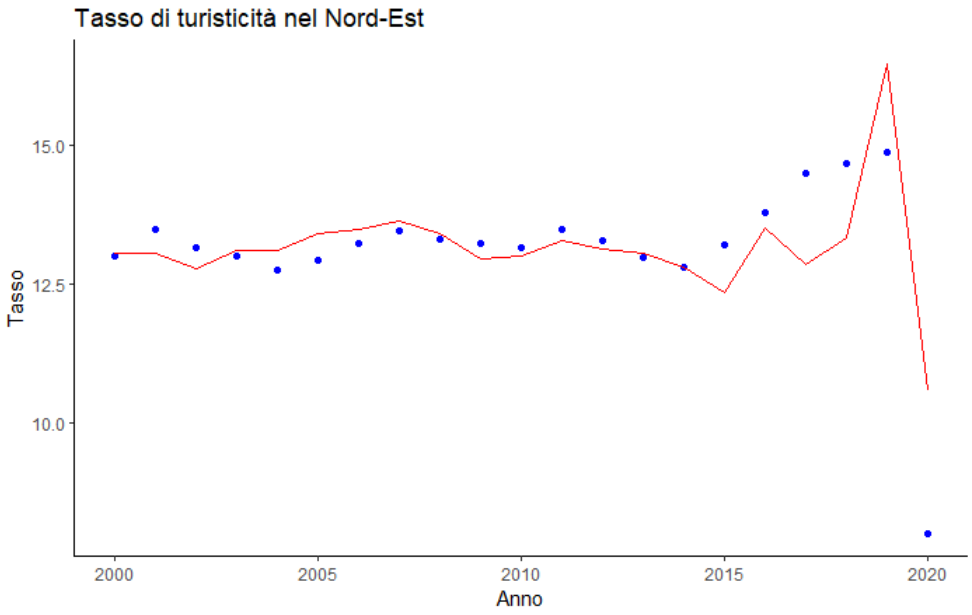
Popolazione

La popolazione è l'unica variabile che riesce a migliorare il modello con i passeggeri e il tempo.



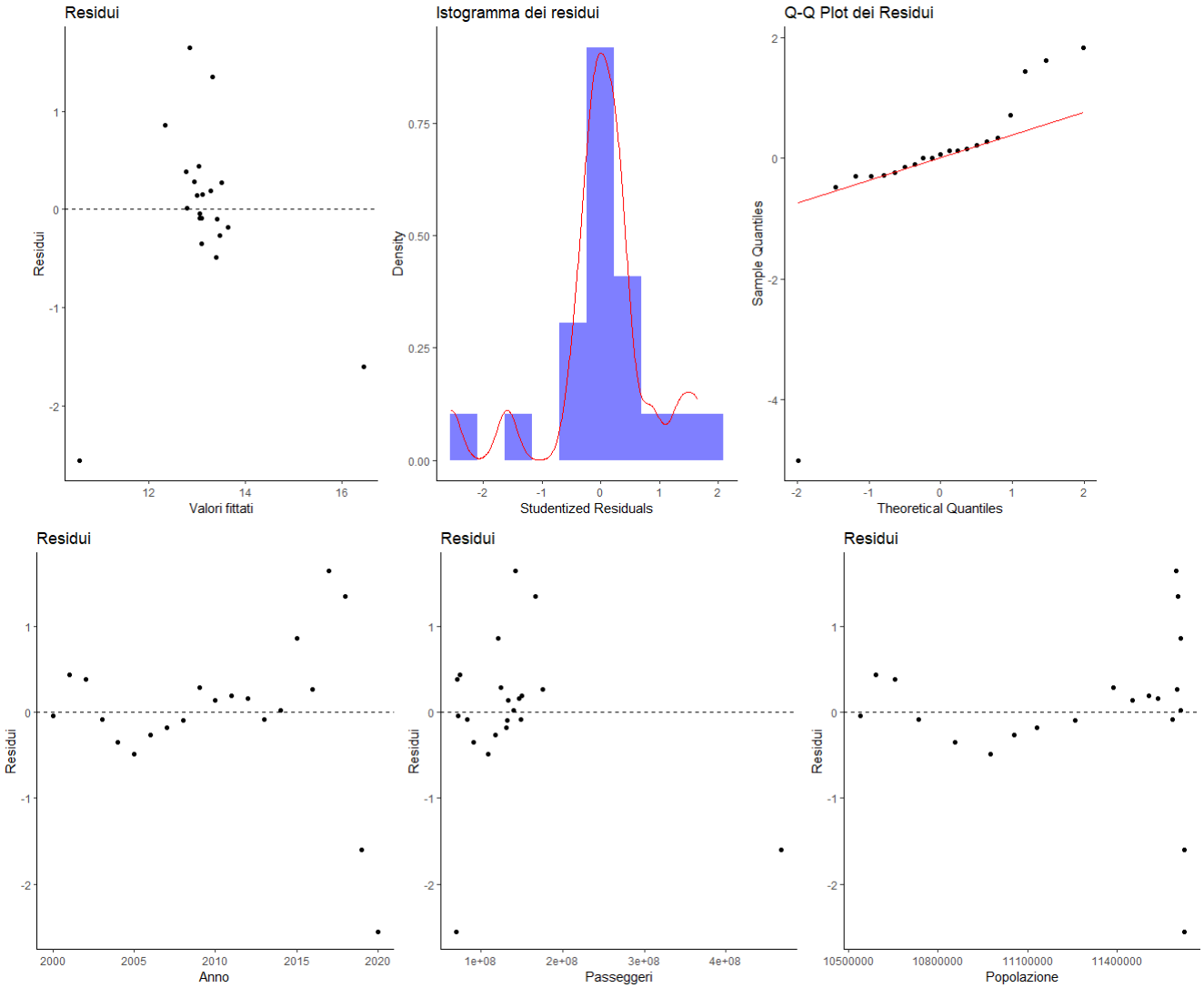
Il risultato che si ottiene aggiungendo la popolazione è piuttosto paradossale, poiché questa non ha minimamente risentito del 2020, ma ha un impatto positivo nel modello nella stima del trend.
IL modello è il seguente[16]:

```
PopPasTimeModel_NE <-lm(turismo ~ anno + log(passeggeri) + popolazione -1, data = NordEst_ts)
```



[16]Summary in tabella 19

Si visualizzano graficamente i residui del modello:



Si osserva graficamente nei residui l'assenza di linearità, omoschedasticità e normalità. I residui risulatno pertanto non accettabili.

Conclusioni

Le conclusioni sulle serie storiche non sono molto felici: non è stato individuato un modello soddisfacente.
Il problema è principalmente che non è stata trovata una variabile in grado di spiegare il dato del 2020, mentre in un modello in cui si va ad escludere l'ultimo dato il tasso diventa quasi stazionario, rendendo lo studio molto semplice con le sole tecniche di serie storica; di fatto, così si vanno a perdere tutte le altre variabili della regressione lineare semplice.

4 Tabelle e altro

```
Call:
lm(formula = turismo ~ movimenti, data = data2000)

Residuals:
   Min     1Q Median     3Q    Max
-7.928 -5.480 -2.201  1.284 29.720

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.914e+00  2.533e+00   3.914  0.00102 **
movimenti   -2.186e-05  2.125e-05  -1.029  0.31719
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.265 on 18 degrees of freedom
Multiple R-squared:  0.05554,Adjusted R-squared:  0.003068
F-statistic: 1.058 on 1 and 18 DF,  p-value: 0.3172
```

Table 1: Summary modello tasso di turisticità e movimenti aerei

```
Call:
lm(formula = newturismo ~ newmovimenti, data = data2000)

Residuals:
    Min      1Q  Median      3Q     Max
-10.511  -2.962  -1.401   3.221  17.458

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.518      3.440  -0.441  0.66540
newmovimenti 1695.628    531.044   3.193  0.00605 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.947 on 15 degrees of freedom
Multiple R-squared:  0.4047,Adjusted R-squared:  0.365
F-statistic:  10.2 on 1 and 15 DF,  p-value: 0.00605
```

Table 2: Summary modello tasso di turisticità e movimenti aerei trasformati

```
Call:
lm(formula = newturismo ~ newmovimenti - 1, data = data2000)

Residuals:
   Min     1Q Median     3Q    Max
-9.980 -3.347 -2.146  2.784 18.781

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
newmovimenti   1491.4      253.5   5.884 2.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.77 on 16 degrees of freedom
Multiple R-squared:  0.6839,Adjusted R-squared:  0.6642
F-statistic: 34.62 on 1 and 16 DF,  p-value: 2.308e-05
```

Table 3: Summary modello tasso di turisticità e movimenti aerei trasformati senza intercetta

```
Call:
lm(formula = newdata$newturismo ~ newdata$newmovimenti - 1, data = data2000)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9424 -1.8595  0.7804  4.0283  9.2360

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
newdata$newmovimenti    989.1      178.8   5.532 5.75e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.078 on 15 degrees of freedom
Multiple R-squared:  0.6711,Adjusted R-squared:  0.6491
F-statistic:  30.6 on 1 and 15 DF,  p-value: 5.754e-05
```

Table 4: Summary modello tasso di turisticità e movimenti aerei trasformati senza intercetta e senza Trentino

```
Call:
lm(formula = turismo ~ produttivita + movimenti, data = data2000)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4664 -3.2754  0.4777  2.3982 12.6811

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.788e+01  7.520e+00  -3.707  0.00175 **
produttivita  8.453e-01  1.642e-01   5.148 8.05e-05 ***
movimenti    -2.806e-05  1.372e-05  -2.045  0.05666 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.96 on 17 degrees of freedom
Multiple R-squared:  0.6309,Adjusted R-squared:  0.5875
F-statistic: 14.53 on 2 and 17 DF,  p-value: 0.0002093
```

Table 5: Summary modello tasso di turisticità, movimenti aerei e produttività

```
Call:
lm(formula = turismo ~ movimenti + produttivita * NS, data = data2000)

Residuals:
    Min      1Q  Median      3Q     Max
-10.209  -1.676  -0.074   3.245   9.476

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -4.881e+01  1.324e+01  -3.687 0.002195 **
movimenti          -2.559e-05  1.270e-05  -2.015 0.062192 .
produttivita        1.251e+00  2.566e-01   4.876 0.000201 ***
NSsud               5.750e+01  2.162e+01   2.660 0.017832 *
produttivita:NSsud -1.369e+00  5.239e-01  -2.613 0.019568 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.23 on 15 degrees of freedom
Multiple R-squared:  0.7492,Adjusted R-squared:  0.6823
F-statistic:  11.2 on 4 and 15 DF,  p-value: 0.0002069
```

Table 6: Summary modello tasso di turisticità, movimenti aerei e produttività con interazione

```
Call:
lm(formula = turismo ~ produttivita + unescoTot + movimenti +
    popolazione + spesa, data = data2000[, -1])

Residuals:
     Min       1Q   Median       3Q      Max
-11.9554  -1.4011   0.5915   1.9427   6.3342

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.535e+01  8.679e+00  -5.225 0.000129 ***
produttivita  1.247e+00  1.808e-01   6.898 7.35e-06 ***
unescoTot    -2.968e+00  8.974e-01  -3.307 0.005185 **
movimenti    -7.182e-05  2.307e-05  -3.113 0.007632 **
popolazione   5.483e-06  1.755e-06   3.125 0.007460 **
spesa        -8.474e-03  3.439e-03  -2.464 0.027304 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.598 on 14 degrees of freedom
Multiple R-squared:  0.8191,Adjusted R-squared:  0.7545
F-statistic: 12.68 on 5 and 14 DF,  p-value: 8.728e-05
```

Table 7: Summary modello tasso di turisticità ottenuto via step-AIC

```
Call:
lm(formula = turismo ~ anno, data = NordEst_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0070 -0.2403  0.0614  0.2681  1.8284

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.54310   97.57988   0.415    0.682
anno        -0.01363    0.04855  -0.281    0.782

Residual standard error: 1.347 on 19 degrees of freedom
Multiple R-squared:  0.004131,Adjusted R-squared:  -0.04828
F-statistic: 0.07882 on 1 and 19 DF,  p-value: 0.7819
```

Table 8: Summary modello Nord-Est con il tempo

```
Call:
lm(formula = turismo ~ anno, data = NordEst_ts)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90167 -0.29229 -0.02248  0.22176  0.82695

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -118.07427   36.73387  -3.214  0.00481 **
anno           0.06543    0.01828   3.579  0.00214 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4714 on 18 degrees of freedom
Multiple R-squared:  0.4158,Adjusted R-squared:  0.3833
F-statistic: 12.81 on 1 and 18 DF,  p-value: 0.002144
```

Table 9: Summary modello Nord-Est con il tempo senza il 2020

```
Call:
lm(formula = 1/sqrt(turismo) ~ anno - 1, data = NordEst_ts)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018908 -0.004121 -0.001997  0.001147  0.074931

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
anno 1.379e-04  1.984e-06   69.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01828 on 20 degrees of freedom
Multiple R-squared:  0.9959,Adjusted R-squared:  0.9957
F-statistic:  4828 on 1 and 20 DF,  p-value: < 2.2e-16
```

Table 10: Summary modello Nord-Est con il tempo e la variabile risposta trasformata

```
Call:
lm(formula = turismo ~ passeggeri, data = NordEst_ts)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60500 -0.18376 -0.09650  0.08943  1.12193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.272e+01  2.015e-01  63.145  < 2e-16 ***
passeggeri  4.904e-09  1.247e-09   3.932 0.000978 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4524 on 18 degrees of freedom
Multiple R-squared:  0.462,Adjusted R-squared:  0.4321
F-statistic: 15.46 on 1 and 18 DF,  p-value: 0.0009778
```

Table 11: Summary modello passeggeri

```
Call:
lm(formula = turismo ~ anno + passeggeri, data = NordEst_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4797 -0.2041  0.0156  0.1929  1.9281

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.978e+02  9.563e+01   2.068  0.05328 .
anno        -9.262e-02  4.770e-02  -1.941  0.06803 .
passeggeri   1.110e-08  3.587e-09   3.094  0.00626 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.118 on 18 degrees of freedom
Multiple R-squared:  0.3499,Adjusted R-squared:  0.2777
F-statistic: 4.844 on 2 and 18 DF,  p-value: 0.02074
```

Table 12: Summary modello tempo passeggeri

```
Call:
lm(formula = turismo ~ movimenti, data = NordEst_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5260 -0.2912 -0.1759  0.1533  1.3300

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.236e+01  4.239e-01  29.146   <2e-16 ***
movimenti   5.813e-07  2.256e-07   2.576    0.019 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5272 on 18 degrees of freedom
Multiple R-squared:  0.2694,Adjusted R-squared:  0.2288
F-statistic: 6.638 on 1 and 18 DF,  p-value: 0.01902
```

Table 13: Summary modello movimenti

```
Call:
lm(formula = turismo ~ anno + movimenti - 1, data = NordEst_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1567 -0.1411 -0.0379  0.5261  1.6542

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
anno      5.456e-03  4.263e-04  12.798 8.67e-11 ***
movimenti 1.237e-06  4.645e-07   2.663   0.0154 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.155 on 19 degrees of freedom
Multiple R-squared:  0.9931,Adjusted R-squared:  0.9924
F-statistic:  1365 on 2 and 19 DF,  p-value: < 2.2e-16
```

Table 14: Summary modello tempo movimenti

```
Call:
lm(formula = turismo ~ anno + log(passeggeri), data = NordEst_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9860 -0.3369 -0.1523  0.4033  1.9082

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     245.71805   76.03834   3.232  0.00463 **
anno             -0.14390    0.04106  -3.504  0.00253 **
log(passeggeri)   3.04277    0.60050   5.067 8.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8885 on 18 degrees of freedom
Multiple R-squared:  0.5896,Adjusted R-squared:  0.544
F-statistic: 12.93 on 2 and 18 DF,  p-value: 0.0003305
```

Table 15: Summary modello tempo passeggeri (logaritmo)

```
Call:
lm(formula = turismo ~ pil, data = NordEst_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2710 -0.0488  0.1003  0.2930  1.4426

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.146e+01  2.910e+00   3.939  0.00088 ***
pil         3.286e-06  5.646e-06   0.582  0.56744
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.338 on 19 degrees of freedom
Multiple R-squared:  0.01751,Adjusted R-squared:  -0.0342
F-statistic: 0.3387 on 1 and 19 DF,  p-value: 0.5674
```

Table 16: Summary modello PIL

```
Call:
lm(formula = turismo ~ anno + log(passeggeri) + pil - 1, data = NordEst_ts)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50797 -0.20663 -0.06542  0.22586  1.88641

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
anno            -2.047e-02  6.082e-03  -3.365 0.003448 **
log(passeggeri)  3.311e+00  7.540e-01   4.392 0.000352 ***
pil             -1.442e-05  5.722e-06  -2.521 0.021372 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9602 on 18 degrees of freedom
Multiple R-squared:  0.9955,Adjusted R-squared:  0.9947
F-statistic:  1319 on 3 and 18 DF,  p-value: < 2.2e-16
```

Table 17: Summary modello passeggeri PIL tempo

```
Call:
lm(formula = turismo ~ spesa, data = NordEst_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1868  0.0082  0.3430  0.4777  1.0190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.397579   2.155909   3.431   0.0028 **
spesa       0.003738   0.001392   2.686   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.149 on 19 degrees of freedom
Multiple R-squared:  0.2752,Adjusted R-squared:  0.237
F-statistic: 7.213 on 1 and 19 DF,  p-value: 0.01463
```

Table 18: Summary modello spesa

```
Call:
lm(formula = turismo ~ anno + log(passeggeri) + popolazione -
    1, data = NordEst_ts)

Residuals:
     Min       1Q   Median       3Q      Max
-2.55314 -0.17812  0.01813  0.28661  1.64684

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
anno            -1.078e-02  4.925e-03  -2.189 0.042022 *
log(passeggeri)  3.095e+00  6.501e-01   4.761 0.000156 ***
popolazione     -2.027e-06  7.056e-07  -2.873 0.010117 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9248 on 18 degrees of freedom
Multiple R-squared:  0.9958,Adjusted R-squared:  0.9951
F-statistic:  1422 on 3 and 18 DF,  p-value: < 2.2e-16
```

Table 19: Summary modello popolazione, passeggeri e anno