Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

Relazione Finale

Influenza dell'inquinamento atmosferico durante la gravidanza sullo sviluppo del bambino: una valutazione basata sull'analisi di dati funzionali

Relatore: Prof. Livio Finos
Dipartimento di Scienze Statistiche

Laureando: Giulia Pigatto
Matricola N. 1198833

Anno Accademico 2021/2022

Indice

Obiettivo della tesi

L'obiettivo di questa tesi consiste nel valutare l'effetto che ha l'inquinamento atmosferico dovuto al PM_{10} sullo sviluppo dei bambini durante la gravidanza. Si studia l'influenza di tale particolato atmosferico a seconda di alcune caratteristiche proprie dei soggetti appartenenti allo studio considerato.

Innanzitutto si vogliono descrivere una ad una tutte le variabili del dataset a disposizione, passando poi all'analisi delle associazioni tra ciascuna variabile e lo sviluppo del bambino. In conclusione si analizzerà un modello di regressione opportuno, evidenziando quali covariate influenzano significativamente lo sviluppo dei soggetti in esame.

Metodi

Le analisi sono state effettuate attraverso il software statistico **R** e la relazione finale è stata redatta tramite l'utilizzo di **R Markdown**.

1 Inquinamento e salute

1.1 L'inquinamento atmosferico

L'inquinamento dell'aria è dato dalla contaminazione dell'ambiente indoor o outdoor da parte di agenti chimici, fisici o biologici che modificano le caratteristiche naturali dell'atmosfera. Apparecchi per il riscaldamento delle abitazioni, i motori dei veicoli, gli impianti industriali e gli incendi boschivi sono comuni sorgenti di inquinamento atmosferico. Il materiale particolato (PM_{10}), il monossido di carbonio (CO), l'ozono (O_3), il biossido di azoto (NO_2) e quello di zolfo (SO_2) sono inquinanti di grande interesse per la salute pubblica.
Questi inquinanti si formano, in parte o interamente, in atmosfera a partire da altre sostanze dette "precursori". La concentrazione osservata di questi agenti inquinanti e la sua variabilità nel tempo e nello spazio dipendono, oltre che dal carico emissivo, anche da altri fattori, legati alla meteorologia e alla reattività chimica delle specie emesse.

L'inquinamento dell'aria danneggia non solo l'ambiente ma anche la salute umana. In Italia, le emissioni di molti inquinanti atmosferici sono diminuite notevolmente negli ultimi decenni, con conseguente miglioramento della qualità dell'aria. Tuttavia, le concentrazioni di inquinanti atmosferici sono ancora troppo elevate e i problemi di qualità dell'aria persistono [1].
Una parte significativa della popolazione europea vive in zone, in particolar modo nelle città, in cui si superano i limiti fissati dalle norme in materia di qualità dell'aria: l'inquinamento da ozono, biossido di azoto e particolato pone gravi rischi per la salute. Circa il 90% degli abitanti delle città è esposto a concentrazioni di inquinanti superiori ai livelli di qualità dell'aria ritenuti dannosi per la salute.
L'ozono, il biossido di azoto e il particolato atmosferico sono attualmente considerati i tre inquinanti che in maniera più significativa incidono sulla salute umana. La gravità dell'impatto delle esposizioni prolungate e di picco a questi inquinanti varia dall'indebolimento del sistema respiratorio fino alla morte prematura [2].

Più recentemente si è visto che l'insieme delle sostanze inquinanti sospese nell'aria che inaliamo non mette in pericolo solamente le vie respiratorie, ma potrebbe avere degli effetti anche sulla salute del cervello. Da quanto emerge da uno studio del Beth Israel Deaconess Medical Center di Boston, affiliato alla Harvard Medical School (Massachusetts), esisterebbe una correlazione tra l'esposizione giornaliera ad alti livelli di inquinamento e una diminuzione del volume del cervello assimilabile a quella causata da un invecchiamento accelerato. Un aumento di 2 $\mu g/m^3$ nella media di particolato fine in atmosfera è associato a una diminuzione dello 0,32% del volume del cervello, la perdita che si registra in un anno di invecchiamento cerebrale [Elissa H. Wilker]. Per il momento si è mostrata solamente

una correlazione tra inquinamento e salute cerebrale, ma non un rapporto di causa-effetto [3].

Per avvicinarsi all'obiettivo della presente tesi, è importante far notare che un altro studio, finanziato dall'Unione Europea, offre prove convincenti sull'impatto di particolari sostanze inquinanti sulla struttura e le funzioni cerebrali, dimostrando una correlazione tra esposizione a inquinamento atmosferico e indebolimento delle prestazioni cognitive nei bambini [4].

1.2 Il PM_{10} come agente inquinante

Il particolato è uno dei maggior responsabili dell'inquinamento atmosferico. L'aria contiene del pulviscolo che può essere innocuo, se di origine naturale e presente in piccole quantità, o dannoso, se abbondante e inalabile. Le fonti possono essere sia di origine naturale che antropica; pertanto, la composizione del particolato può risultare molto varia. Si possono distinguere due classi principali di questo agente inquinante, divise sia per composizione sia per dimensioni: particolato grossolano e particolato fine.
Il particolato grossolano è costituito da particelle, compresi pollini e spore, con diametro superiore a 10 μm. Sono in genere trattenuti dalla parte superiore dell'apparato respiratorio (naso e laringe). Vengono definite polveri fini le particelle di polvere con un diametro aerodinamico inferiore a 10 μm (PM_{10}), in grado di penetrare nel tratto respiratorio superiore (naso, faringe e trachea) e le particelle con diametro inferiore a 2,5 μm ($PM_{2,5}$), particolato fine in grado di penetrare profondamente nei polmoni specie durante la respirazione dalla bocca.

Il particolato aerodisperso è in grado di adsorbire gas e vapori tossici sulla superficie delle particelle. Tale fenomeno contribuisce ad aumentare le concentrazioni degli inquinanti gassosi che raggiungono le zone più profonde del polmone, trasportati dalle particelle PM_{10} e $PM_{2,5}$.
Numerosi studi [1] hanno evidenziato una correlazione tra esposizione acuta a particolato aerodisperso e sintomi respiratori, alterazioni della funzionalità respiratoria, ricoveri in ospedale e mortalità per malattie respiratorie. Inoltre, l'esposizione prolungata nel tempo a particolato, già a partire da basse dosi, è associata all'incremento di mortalità per malattie respiratorie e di patologie quali bronchiti croniche, asma e riduzione della funzionalità respiratoria. L'esposizione cronica, inoltre, è verosimilmente associata a un incremento di rischio di tumore delle vie respiratorie [5].

Anche se l'inquinamento da particolato colpisce tipicamente gli apparati respiratorio e cardiovascolare, alcuni studi hanno evidenziato l'associazione tra inquinamento, in parti-

colare quello da PM_{10}, e lo sviluppo neurologico dei bambini esposti a questo inquinante. A questo proposito, alcuni studi epidemiologici avevano cercato di spiegare l'inquinamento atmosferico come causa dei disturbi dello sviluppo neurologico, ma mancavano di prove definitive di questo collegamento.
Uno studio condotto nel 2017, invece, il quale si avvale di una coorte basata sulla popolazione estratta dal database National Health Insurance Service (NHIS) della Corea, ha individuato una correlazione tra l'esposizione al particolato PM_{10} e al biossido di azoto (NO_2) e l'insorgenza del disturbo da deficit di attenzione e iperattività nell'infanzia. In questo studio l'ADHD (Disturbo da Deficit di Attenzione e Iperattività) infantile si è verificato nel 3,5% dei soggetti e, con l'aumento di 1 $\mu g/m^3$ di inquinanti atmosferici, gli HR dell'ADHD infantile erano 1,18 (IC 95%: 1,15-1,21) in caso di PM_{10} e 1,03 (IC 95%: 1,02-1,04) in caso di NO_2. Dunque questo disturbo è stato associato all'esposizione a PM_{10} e NO_2.

Nell'aprile 2008 l'Unione Europea ha adottato definitivamente una nuova direttiva (2008/50/EC) che detta limiti di qualità dell'aria con riferimento anche al $PM_{2,5}$. Tale direttiva è stata recepita dalla legislazione italiana con il Decreto Legislativo 155/2010, il quale stabilisce le soglie di concentrazione in aria delle polveri sottili PM_{10}, calcolate su base temporale giornaliera e annuale [6]. Si riportano in Tabella 1.1 i valori limite delle concentrazioni giornaliere e annuali di PM_{10} e il numero massimo previsto di superamenti, stabiliti in Italia e in Europa ma anche in alcuni altri paesi nel mondo.

	Valore massimo per la media annuale	Valore massimo giornaliero (24 ore)	Numero massimo superamenti consentiti in un anno
Italia e Europa	40	50	35
Australia	-	50	-
Cina	70	150	-
Giappone	-	100-200	-
Russia	40	60	-
USA	-	150	1
OMS (2005)	20	50	-

Tabella 1.1: Valori di soglia massima delle concentrazioni giornaliere e annuali di PM10 riportati in $\mu g/m^3$ e il numero massimo previsto di superamenti, stabiliti in Italia ed Europa ma anche in alcuni altri paesi nel mondo.

In Italia, le emissioni di PM_{10} sono caratterizzate nel periodo 1990 – 2012 da un andamento decrescente, passando da 239 Gg a 153 Gg con un decremento del 36% (-16% dal 2003). La maggior parte delle emissioni è dovuta alla combustione non industriale (riscaldamento,

41% del totale nel 2012) e ai trasporti su strada (17% nel 2012). Le altre sorgenti mobili, pesando per il 9% delle emissioni nazionali, mostrano una riduzione di circa il 58%. Importante sottolineare l'andamento crescente delle emissioni da riscaldamento per le quali si registra un incremento del 113% rispetto all'anno base. La riduzione più evidente (-94%) si riscontra nelle emissioni derivanti dalla combustione per la produzione di energia e nelle industrie di trasformazione, il cui contributo pari al 19% nel 1990 risulta inferiore al 2% nel 2012 [7].

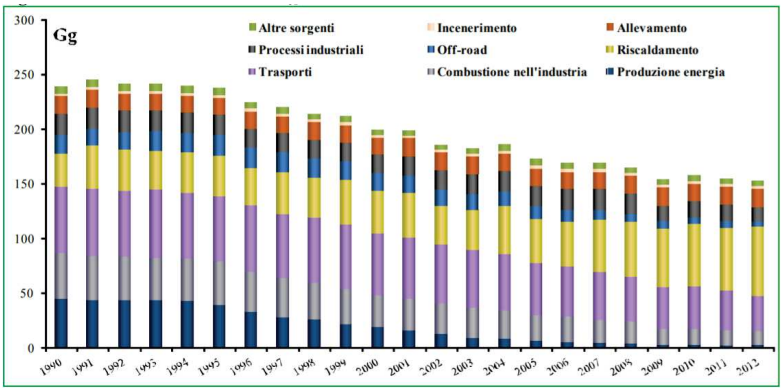

Figura 1.1: Emissioni di PM_{10} in Italia dal 1990 al 2012.

1.3 Effetti del PM_{10} durante la gravidanza

L'esposizione agli inquinanti atmosferici durante la gestazione è stata recentemente identificata come un potenziale fattore di rischio per alcuni danni neuropsicologici, in particolare l'inquinamento atmosferico sembra avere degli effetti sullo sviluppo cognitivo e psicomotorio dei bambini di età compresa tra 1 e 6 anni.
Lo sviluppo del cervello e la crescita degli organi avvengono in fasi che iniziano nel periodo prenatale. Gli agenti atmosferici possono influenzare lo sviluppo neurologico in modo diverso a seconda dei tempi di esposizione e del sesso fetale. L'inquinamento dell'aria, specialmente il particolato e i suoi componenti solubili, si deposita sugli alveoli polmonari e, attraverso l'apparato circolatorio, può raggiungere gli organi come il cervello.
Uno studio conseguito su sei coorti di nascita basate sulla popolazione europea, tra le quali si colloca GASPII in Italia, ha reclutato coppie madre-figlio dal 1997 al 2008 estrapolando i dati sul livello di inquinamento ai periodi della gravidanza. Da questo studio è emerso che esiste un'associazione tra l'esposizione al particolato atmosferico (PM_{10} e $PM_{2,5}$) e all'NO_2 e una riduzione dello sviluppo psicomotorio durante l'infanzia.

Il primo trimestre di gravidanza si caratterizza per l'avvio del processo di organogenesi, ossia il processo di generazione degli organi quali cuore, stomaco, organi del sistema nervoso,

e per la comparsa e l'allungamento degli arti. In particolare, il sistema nervoso inizia il suo sviluppo intorno alla terza settimana dal concepimento, quando l'ectoderma della placca neurale comincia a differenziarsi in tessuto nervoso. Nella prima parte della gravidanza è particolarmente evidente l'accrescimento del cervello, più veloce di quello corporeo. Il periodo critico per lo sviluppo cerebrale si colloca tra la dodicesima e le sedicesima settimana di gravidanza, in cui ha inizio una grande proliferazione e migrazione dei neuroni, che dà origine alle diverse aree del cervello.
Nelle ultime dieci settimane il notevole accrescimento del cervello, la differenziazione citoarchitettonica e lo sviluppo concorrente delle funzioni aumentano le richieste di ossigeno e, quindi, la vulnerabilità all'ipossia. Le conseguenze di questa condizione in questo stadio possono essere responsabili di deficit cognitivi.
Per questo motivo, in questa tesi verrà analizzato come variano le emissioni del particolato PM_{10} nell'ambiente in cui vive la madre per ciascun trimestre o mese della gravidanza, così da poter comprendere, ad esempio, se alte concentrazioni di PM_{10} nel periodo in cui si stabiliscono le funzioni del cervello possano influire sullo sviluppo cerebrale e cognitivo del bambino in modo più rilevante rispetto ad altri momenti della gestazione [8].

2 Analisi descrittive

In questo capitolo viene presentata la prima parte dell'analisi, dedicata alla descrizione del dataset in ogni singola variabile e alla sintesi dei dati. Il dataset iniziale analizzato contiene i dati relativi a 2776 bambini, di età compresa tra le 2 settimane e gli 8 anni. Per questa tesi si è voluta focalizzare l'attenzione sui bambini di età uguale o superiore ai 3 anni.
I soggetti considerati sono nati tra il 2006 e il 2011 e provengono da tutta Italia. Per ciascuno di essi sono state rilevate numerose variabili, descritte in seguito, e in particolare a ciascun bambino sono state associate le concentrazioni di PM_{10} presenti durante la gravidanza sulla base della loro città di residenza.
Le variabili di interesse in questo studio sono gli indici di sviluppo cognitivo, motorio e comportamentale, riferiti ai bambini oggetto di studio.
Come per i soggetti, anche per quanto riguarda le variabili si è dovuto fare una selezione in modo da considerare solo quelle caratteristiche che possano influenzare maggiormente l'analisi.

2.1 Descrizione delle variabili

Il dataset selezionato si compone di 1669 soggetti e 40 variabili. Dato che descrivere ogni singola caratteristica sarebbe poco comprensibile, oltre che dispendioso, si espongono le variabili in gruppi pressoché tematici.

- *anno, mese, eta_in_mesi, eta_cat, genere, rip_geo, regione, naz_bambino*: sono delle caratteristiche demografiche e di residenza che riguardano il bambino. Anno e mese di nascita, l'età (in mesi) espressa sia come numerica sia come categoriale, e una serie di variabili categoriali come il sesso, il luogo di nascita (ripartizione geografica e regione) e la nazionalità.
- *motorio, adattivo, socioemotivo, cognitivo, comunicativo, totale*: sono gli indici di sviluppo del bambino oggetto di studio. Essi sono espressi da un valore numerico discreto.
- *pm10_nascita, pm10_gravid, pm10_1m_pre, pm10_2m_pre, pm10_3m_pre, pm10_4m_pre, pm10_5m_pre, pm10_6m_pre, pm10_7m_pre, pm10_8m_pre*: sono la media delle concentrazioni di PM_{10} registrate durante l'intera gravidanza e la media per ciascun mese della stessa. La variabile *pm10_1m_pre*, ad esempio, indica la concentrazione media di PM_{10} registrata 1 mese prima della nascita, mentre *pm10_nascita* è quella al momento della nascita. Esse sono variabili di tipo numerico.
- *mstatocivile, pstatocivile, naz_madre, naz_padre, naz_genitori, titolo_max, lavoro_padre, lavoro_madre, n_figli, n_figli_cat, fratelli, minore, maggiore*: è un gruppo di variabili riferite alla situazione familiare e alla formazione dei genitori dei

soggetti in esame. Le variabili riguardanti lo stato civile, la nazionalità, il titolo massimo conseguito e l'occupazione dei genitori sono categoriali; le variabili fratelli, minore e maggiore, che indicano rispettivamente se ci siano altri figli oltre al bambino oggetto di studio, e se il soggetto abbia dei fratelli minori o maggiori, sono invece dicotomiche. Le variabili *n_figli* e *n_figli_cat* esprimono entrambe il numero di figli: la prima è di tipo numerico, mentre la seconda è categoriale.

- *interv, tipo, luogo*: sono delle variabili che fanno riferimento alla raccolta dei dati. Esse sono di tipo categoriale e indicano rispettivamente chi ha fatto l'intervista e raccolto i dati, il tipo di questionario somministrato e il luogo di raccolta del questionario.

2.2 Pulizia del dataset

Il dataset studiato per svolgere tutte le analisi in seguito presenta alcuni dati mancanti. Per poter risolvere il problema è possibile intraprendere diverse strade, a seconda della rilevanza delle variabili nelle analisi e dal tipo di studio.
La matrice dei dati iniziale constatava di 100 variabili. Un gruppo di queste, ad esempio, riguardava l'indirizzo di residenza, pertanto si è scelto di mantenere solamente le variabili che indicano la ripartizione geografica e la regione di appartenza dei soggetti in esame.
Quelle variabili che non sono rilevanti ai fini dell'analisi e per le quali si ha un numero considerevole di valori mancanti, non sono state mantenute nel dataset. Altre variabili, invece, risultano avere alcuni valori mancanti ma che sono utili ai fini delle analisi statistiche. In questo caso si possono eliminare tutte le unità per cui mancano questi valori, altrimenti si sostituisce il dato mancante con un indice di posizione della variabile in questione (solamente nel caso in cui essa sia numerica), come la mediana, la media o la moda, o ancora con un altro valore opportunatamente costruito.

Avendo a disposizione un numero considerevole di unità statistiche, quelle che presentano valori mancanti sono state tolte dalla matrice dei dati. Inoltre si elimina dal dataset l'unità statistica per cui risulta che la madre è vedova, essendo l'unico soggetto con questa modalità.
In conclusione, le variabili del dataset finale sono le 40 presentate al paragrafo precedente e, come si evince dalla Tabella 2.2, dei 1669 soggetti del dataset ne verranno inclusi 1468 nelle analisi successive.

Variabile	N	%
titolo_max	5	0.003
lavoro_padre	19	0.011
lavoro_madre	2	0.001
luogo	178	0.107
mstatocivile='vedova'	1	0.001
soggetti inclusi	1468	88

Tabella 2.1: Numero di dati mancanti per ogni variabile e rispettiva percentuale sul totale.

2.3 Analisi univariate

In questo paragrafo viene effettuata un'analisi preliminare univariata delle variabili a disposizione nello studio. Avendo un numero consistente di variabili, esse verranno esposte a seconda del gruppo tematico a cui appartengono.

Dei 1468 soggetti nel campione considerato, 759 sono di genere maschile e 709 sono di genere femminile. Come si può notare dalla Figura 2.1, la maggior parte di essi sono di cittadinanza italiana: solamente 50 bambini sono stranieri. Vi è quindi una predominanza di bambini italiani, e questo probabilmente è dovuto al fatto che i dati sono stati raccolti principalmente presso luoghi dove vi è una prevalenza di cittadini italiani (in riferimento alla Figura 2.11 in seguito).

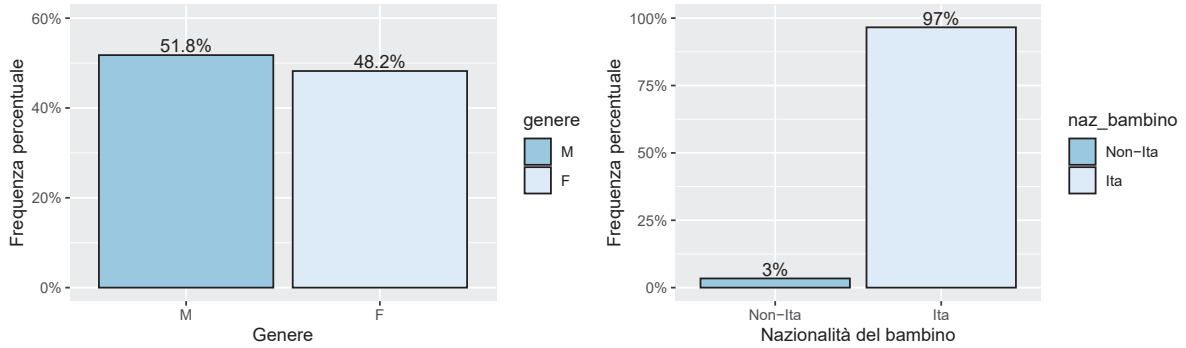

Figura 2.1: Distribuzioni percentuali delle variabili sesso e nazionalità dei soggetti dello studio.

Come si può osservare dalle statistiche descrittive in Tabella 2.2, i soggetti dello studio hanno un'età che varia tra i 36 e i 93 mesi, ovvero hanno all'incirca dai 3 agli 8 anni. Dal test di Shapiro-Wilk si riscontra che la variabile età non ha una distribuzione normale, infatti si rifiuta l'ipotesi nulla di normalità con un p-value inferiore all'1%. Inoltre in Figura 2.2 è riportato il grafico a barre della variabile categoriale riferita all'età, suddivisa in 10 classi di uguale ampiezza. Si può notare che l'età ha una distribuzione asimmetrica a destra, con un valore mediano pari circa a 73 mesi (6 anni), e il 50% della popolazione "centrale" si colloca tra i 57 e gli 80 mesi (4,5 - 6,5 anni).

Variabile	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Età del bambino (in mesi)	36	57.25	72.75	68.64	80.18	93.08	14.5
Anno di nascita	2006	2007	2008	2008	2009	2011	1.3

Tabella 2.2: Statistiche descrittive per le variabili relative all'età del bambino e all'anno di nascita.

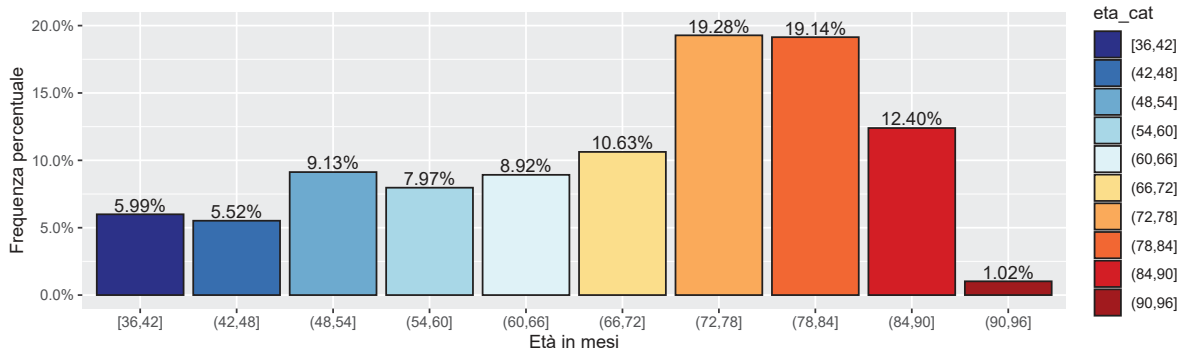

Figura 2.2: Distribuzioni percentuali dell'età dei soggetti dello studio, suddivisa in classi di età.

Per quanto riguarda la provenienza geografica dei soggetti appartenenti allo studio in questione, il 72% è residente al Nord Italia, in particolare approssimativamente la metà della popolazione campionaria proviene dal Veneto, mentre circa il 15% è residente al Sud e Isole e il restante 13% vive al Centro Italia. Dalla Figura 2.3 si può osservare, oltre al grafico a barre della ripartizione geografica, anche quello relativo alle regioni di appartenenza dei soggetti: dopo il Veneto, le regioni da cui provengono i bambini sono Lombardia (17,57%), Sicilia (6,95%), Toscana (6,74%), Trentino-Alto Adige (4,84%), Sardegna (4,77%) e altre regioni per circa il 13% del campione.

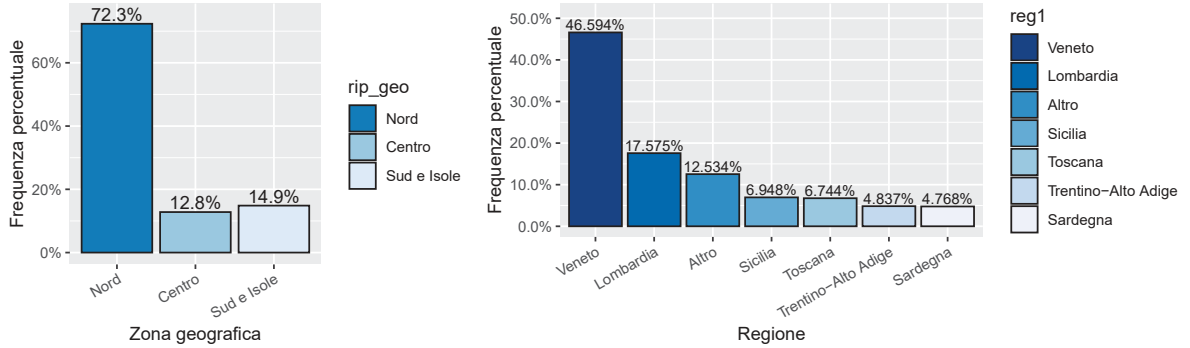

Figura 2.3: Distribuzioni percentuali delle variabili relative alla ripartizione geografica e alla regione di appartenenza dei soggetti dello studio.

Di seguito sono presentate le variabili riguardanti gli indici di sviluppo dei bambini nello studio. Ciascuno di essi è una scala di misurazione ordinale con un range specifico di valori. Essi si suddividono in motorio, adattivo, socioemotivo, cognitivo, comunicativo e totale. Quest'ultimo, in particolare, risulta essere una somma degli indici appena citati, così da fornire un unico valore di sintesi per lo sviluppo di un soggetto.
A partire dai boxplot in Figura 2.4 si può osservare che gli indici di sviluppo presentano diversi outliers sulla coda sinistra della distribuzione: questo significa che vi sono alcuni

soggetti con indici di sviluppo particolarmente bassi. Le distribuzioni di questi indici presentano un'asimmetria negativa.
L'indice di sviluppo motorio comprende i valori tra 11 e 35, con un valore mediano pari a 32. L'indice di sviluppo adattivo, invece, è compreso tra 7 e 37, con un valore mediano pari a 27. Gli indici di sviluppo socioemotivo, cognitivo e comunicativo hanno rispettivamente range di valori (12-36), (12-38) e (11-34) con valori mediani pari a 27, 29 e 26. Gli scarti quadratici medi sono valori indicativamente intorno a 4.
Per quanto riguarda invece l'indice di sviluppo totale, esso comprende valori tra 65 e 175, con un valore mediano pari a 142 e uno scarto quadratico medio di 18. La distribuzione presenta un'asimmetria negativa dovuta a numerosi outliers sulla coda sinistra.
Le distribuzioni delle variabili relative agli indici di sviluppo presentano molti valori anomali e sembrano discostarsi dall'andamento di una normale, infatti il test di Shapiro-Wilk conferma questa ipotesi fornendo un p-value inferiore all'1%.

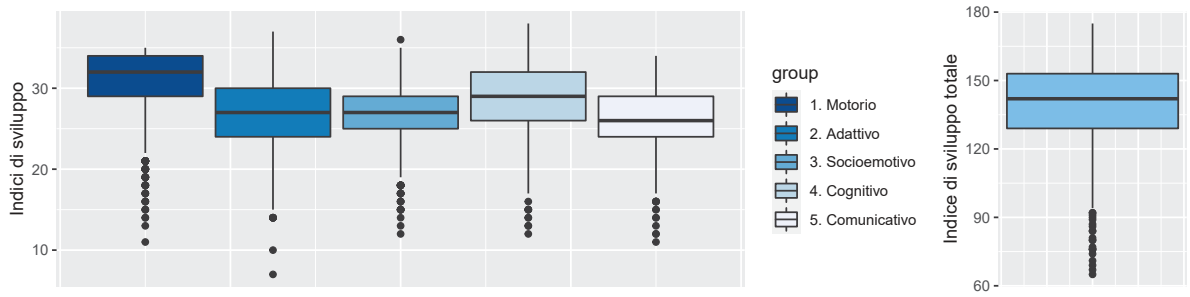

Figura 2.4: Distribuzioni degli indici di sviluppo dei soggetti.

Per lo studio considerato nella presenti tesi si sono raccolti i dati relativi alle concentrazioni di PM_{10} durante la gravidanza delle madri dei soggetti. Nello specifico, si sono rilevate le concentrazioni medie mensili per ciascuno dei mesi della gestazione (da 8 mesi prima della nascita fino alla nascita), e la concentrazione media di PM_{10} durante l'intero periodo di gravidanza.
In Figura 2.5 sono riportati i boxplot delle concentrazioni di emissioni di PM_{10} ordinati nel tempo per ciascun mese, affiancati dal boxplot delle emissioni di particolato nell'intero periodo. Le medie mensili delle concentrazioni di PM_{10} hanno una distribuzione che presenta una forte asimmetria negativa, in effetti il test di Shapiro-Wilk porta al rifiuto dell'ipotesi di normalità per tutte le distribuzioni con un p-value osservato inferiore all'1%. Si può osservare che, in riferimento ai boxplot di sinistra, vi sono numerosi valori anomali in corrispondenza dei valori maggiori.
Nel grafico sono evidenziate due linee tratteggiate in corrispondenza di due soglie specifiche, pari a 40 $\mu g/m^3$ e 50 $\mu g/m^3$: sono rispettivamente i valori massimi consentiti per la media annuale e giornaliera in Italia e in Europa (si veda la Tabella 1.1).

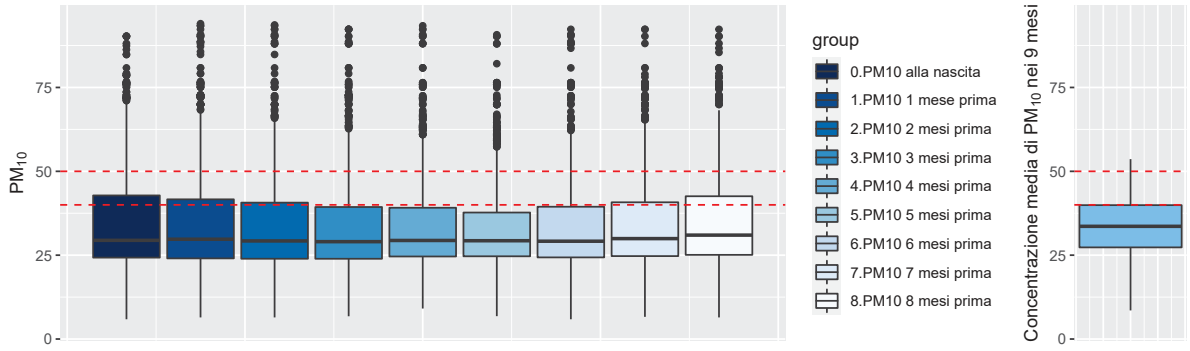

Figura 2.5: Distribuzioni delle concentrazioni medie di PM_{10} in ciascun mese di gravidanza e della concentrazione media di PM_{10} durante tutto il periodo di 9 mesi.

E' interessante osservare l'andamento delle concentrazioni di PM_{10} per ciascuna ripartizione geografica e per ogni regione italiana, poiché è noto che l'inquinamento è presente molto di più al Nord Italia rispetto al Sud e Isole. Per analizzare il confronto tra queste distribuzioni si veda il paragrafo 2.4.

Si procede l'analisi focalizzando l'attenzione sulle variabili relative ai genitori dei soggetti appartenenti allo studio. Nella Figura 2.6 vengono esposti i diagrammi a barre dello stato civile della madre e del padre: vi è una netta evidenza che i genitori siano per lo più sposati o conviventi, infatti rappresentano all'incirca il 93% del campione considerato.

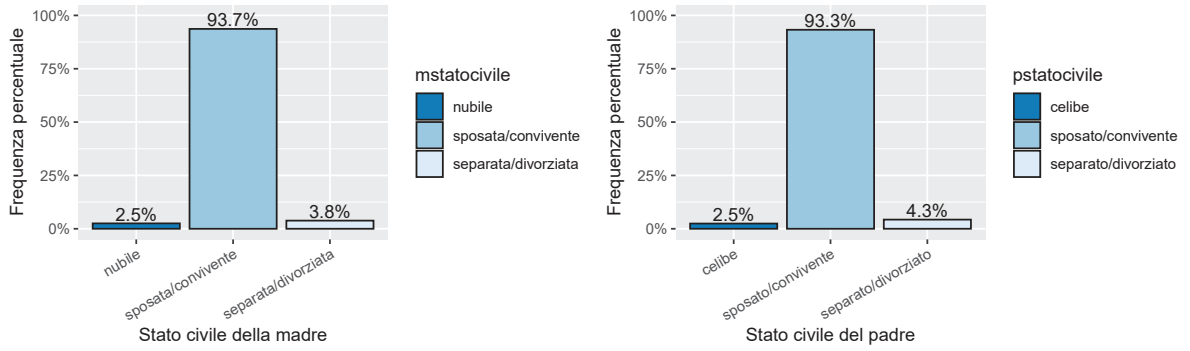

Figura 2.6: Distribuzioni percentuali delle variabili relative allo stato civile della madre e del padre dei soggetti dello studio.

Per quanto riguarda, invece, la nazionalità dei genitori dei bambini nello studio, si ha che il 94% delle madri sono italiane e il 97% dei padri sono italiani. Nello specifico, i soggetti hanno per il 93,3% entrambi i genitori italiani, il 3,3% hanno solo il padre italiano, lo 0,8%

hanno solo la madre italiana, mentre il restante 2,6% ha entrambi i genitori stranieri. Dalla Figura 2.7 si possono osservare i grafici a barre delle variabili relative alla nazionalità dei genitori.

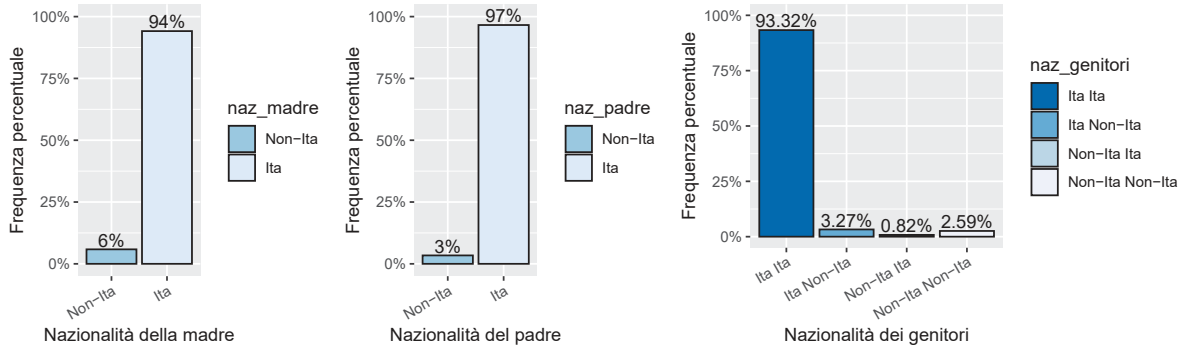

Figura 2.7: Distribuzioni percentuali delle variabili relative alla nazionalità dei genitori dei soggetti dello studio.

Altre variabili interessanti ai fini dell'analisi sono quelle relative al titolo di studio e al lavoro dei genitori. Dalla Figura 2.8 si può notare che metà di essi hanno conseguito un titolo di medio livello, ovvero hanno un diploma, il 37% ha un alto livello di studio, quindi hanno almeno una laurea, e il restante 12% ha un basso titolo di studio, che comprende la licenza elementare o media.
Per quanto riguarda l'occupazione lavorativa, invece, emerge una notevole differenza tra padri e madri: si ha che tra i primi il 92% ha un lavoro full-time, mentre tra le seconde solamente il 40% ha un contratto da full-time. Tra le madri, infatti, il 32% ha un lavoro part-time, il 24% è casalinga e il 5% è disoccupata o frequenta gli studi. I disoccupati o studenti tra i padri sono invece meno del 4%, e chi svolge un lavoro a tempo parziale è circa il 5%.

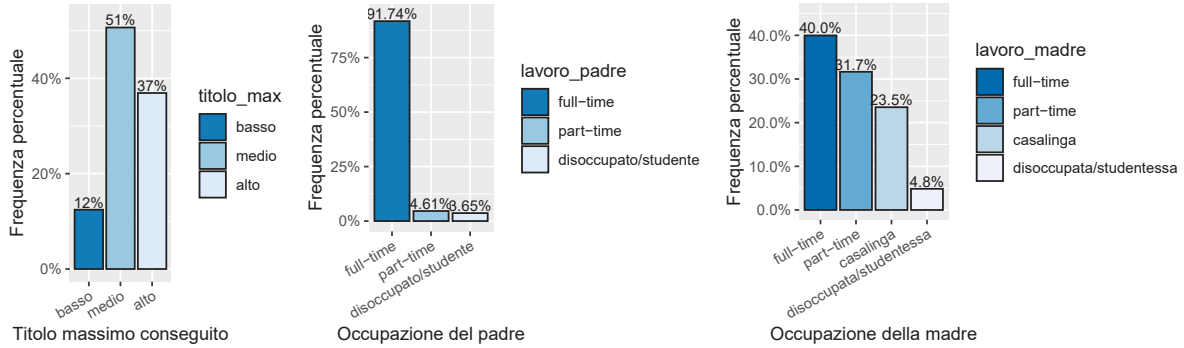

Figura 2.8: Distribuzioni percentuali delle variabili relative al titolo massimo conseguito e all'occupazione svolta dai genitori dei soggetti dello studio.

Nella Figura 2.9 si analizza la variabile relativa al numero di figli. Nel grafico di sinistra si è considerata una variabile categoriale a 3 livelli, che esclude il soggetto nel conteggio. Nei grafici a destra, invece, si osserva la distribuzione del numero di figli come variabile discreta, che comprende il soggetto dello studio nel conteggio.
Riassumendo, si può notare che quest'ultima variabile è unimodale con valore più frequente pari a 2: più di metà (55%) dei soggetti ha un fratello o una sorella. Il 27% dei bambini nel campione è figlio/a unico/a e il 18% ha 2 o più fratelli.

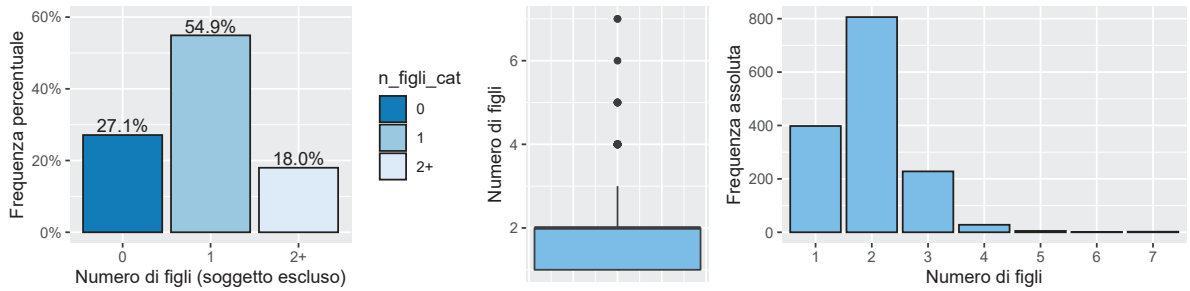

Figura 2.9: Distribuzione della variabile relativa al numero di figli. In particolare nel primo grafico i soggetti considerati nello studio sono esclusi, mentre negli altri due grafici i soggetti sono inclusi.

In Figura 2.10 si rappresentano le variabili dicotomiche che indicano se i soggetti considerati nello studio hanno o meno dei fratelli, e in particolare se hanno fratelli maggiori o minori. Si può osservare che il 73% dei bambini del campione ha almeno un fratello (quindi il 27% è unico/a figlio/a), mentre il 56% ha almeno un fratello maggiore e il 60% ha almeno un fratello minore.

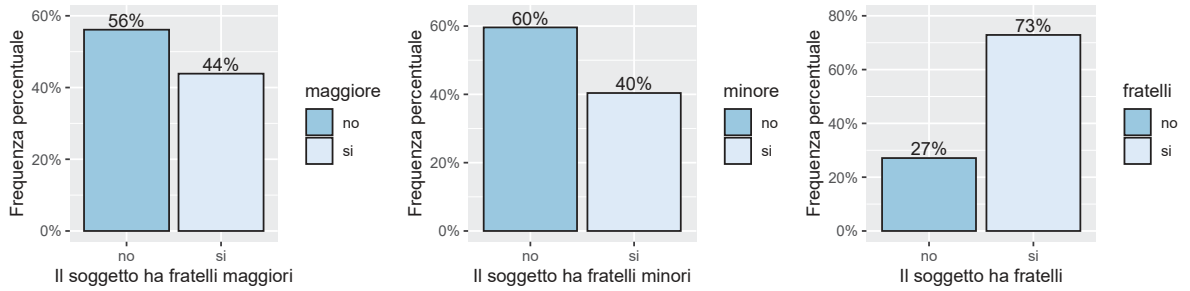

Figura 2.10: Distribuzioni percentuali delle variabili che indicano se i soggetti nello studio hanno fratelli maggiori, fratelli minori o fratelli in generale.

In seguito vengono analizzate alcune variabili riguardanti la raccolta dei dati. Come si nota dalla Figura 2.11, nel 52% dei casi sono state effettuate delle interviste, mentre il 48% dei dati sono stati rilevati tramite un questionario. Dunque, la raccolta dei dati è avvenuta in modo bilanciato tra interviste e questionari.

Per quanto riguarda il luogo di raccolta, invece, si ha che all'incirca il 56% delle rilevazioni sono state effettuate presso le scuole, il 28% dei dati sono stati raccolti nelle abitazioni dei soggetti, nel 12% dei casi attraverso conoscenze personali e il 4% presso una sede sportiva.

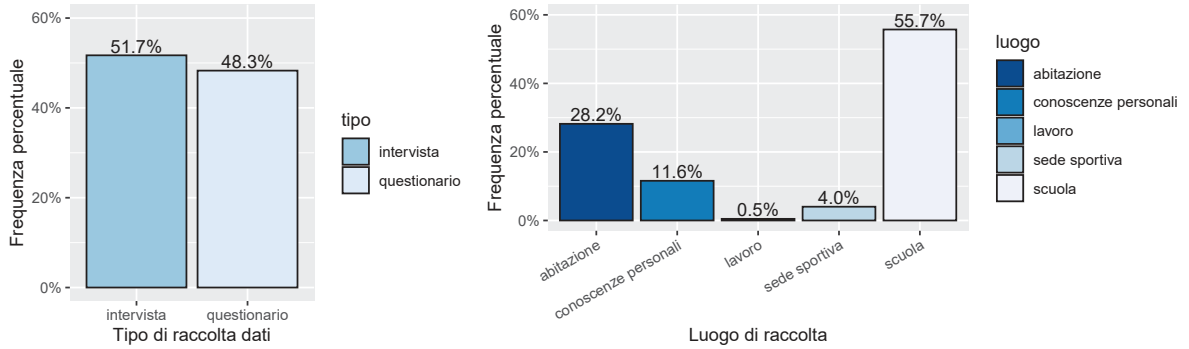

Figura 2.11: Distribuzioni percentuali delle variabili relative al tipo di rilevazione e al luogo di raccolta dei dati.

Per questo studio 29 persone hanno contribuito a raccogliere i dati, in particolare gli intervistatori sono riportati nella Tabella 2.3 e a ciascuno di essi è associato il numero assoluto e percentuale di rilevazioni effettuate.

Variabile	Freq. assoluta	Freq. rel. %	Variabile	Freq. assoluta	Freq. rel. %
rilevatore 1	131	8.9	rilevatore 16	41	2.79
rilevatore 2	121	8.24	rilevatore 17	40	2.72
rilevatore 3	102	6.94	rilevatore 18	36	2.45
rilevatore 4	99	6.74	rilevatore 19	30	2.04
rilevatore 5	97	6.6	rilevatore 20	30	2.04
rilevatore 6	89	6.06	rilevatore 21	30	2.04
rilevatore 7	79	5.38	rilevatore 22	30	1.97
rilevatore 8	75	5.11	rilevatore 23	29	0.82
rilevatore 9	73	4.97	rilevatore 24	12	0.61
rilevatore 10	62	4.22	rilevatore 25	9	0.54
rilevatore 11	54	3.68	rilevatore 26	8	0.54
rilevatore 12	52	3.54	rilevatore 27	7	0.48
rilevatore 13	49	3.34	rilevatore 28	6	0.41
rilevatore 14	47	3.2	rilevatore 29	5	0.34
rilevatore 15	47	3.2			

Tabella 2.3: Distribuzioni di frequenza assoluta e relativa percentuale della variabile relativa all'intervistatore che ha raccolto i dati.

2.4 Analisi bivariate

In questo paragrafo vengono riportate le analisi descrittive bivariate tra le variabili oggetto di studio, ovvero gli indici di sviluppo, e le altre variabili presenti nel dataset.
Tuttavia, inizialmente si vuole confrontare l'andamento delle concentrazioni di PM_{10} per ripartizione geografica e per alcune regioni italiane. Nella Figura 2.12 si possono osservare i diversi andamenti, in particolare nel grafico di destra le concentrazioni di PM_{10} si riferiscono alle tre regioni per le quali si dispongono più informazioni.

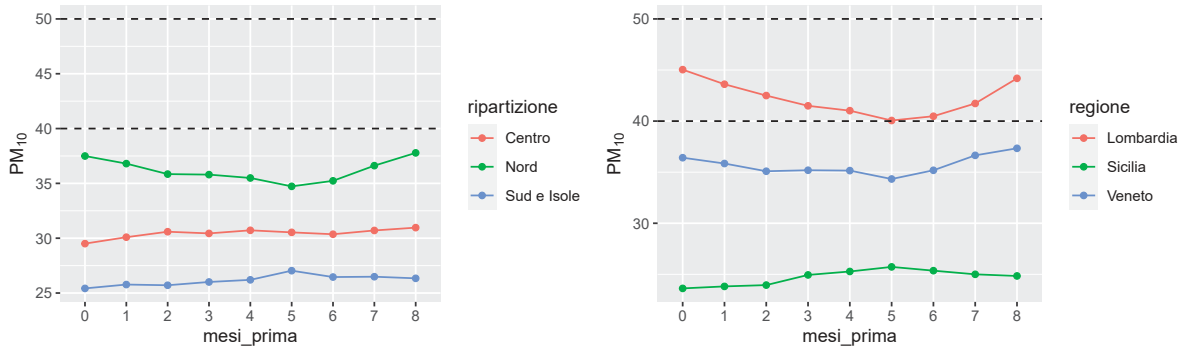

Figura 2.12: Andamento delle concentrazioni medie di PM_{10} durante la gravidanza per ripartizione geografica e per le regioni Veneto, Lombardia e Sicilia.

Prima di esporre le analisi bivariate, è opportuno osservare che le condizioni necessarie per poter sfruttare un approccio parametrico non sono verificate, poiché le variabili relative agli indici di sviluppo non hanno una distribuzione normale, nemmeno per ogni gruppo che si vuole confrontare. Questo risultato è stato ottenuto tramite il test di Shapiro-Wilk.
Per verificare se esiste o meno una differenza significativa di ciascun indice di sviluppo a seconda delle variabili esplicative presentate fino a questo momento, sono stati utilizzati diversi test in base alla natura delle variabili.
Per le variabili categoriali a due livelli, come *genere, children_nationality, tipo*, l'ipotesi di uguaglianza in media tra i gruppi è stata verificata attraverso il test non parametrico di Mann-Whitney. Per quanto riguarda le variabili categoriali a più livelli, come *rip_geo, regione, mstatocivile, pstatocivile, parents_nationality, max_degree, job_father, job_mother, n_children_cat, luogo*, si è utilizzato il test non parametrico di Kruskal-Wallis e in seguito si è svolta un'analisi post-hoc non parametrica tramite i confronti multipli con il metodo di Holm.

Si analizza ora la relazione tra ciascuno degli indici di sviluppo e le caratteristiche demografiche dei soggetti nello studio. Per maggiore facilità di lettura, in seguito vengono analizzati solamente i confronti più significativi tramite boxplot, mentre i restanti sono riportati in Appendice.

In un primo momento si è studiato come si distribuiscono gli indici di sviluppo a seconda del genere e della nazionalità del bambino: a partire dalla Figura 2.13 si osserva che gli indici di sviluppo non sembrano differenziarsi di molto tra maschi e femmine, con valori mediani quasi coincidenti e variabilità molto simili, tuttavia si ha che il 50% dei valori centrali degli indici per le femmine variano su intervalli un po' più elevati.
In ciascun grafico è riportato anche il p-value associato al test di Mann-Whitney per la verifica non parametrica dell'uguaglianza tra le mediane dei due gruppi, individuati dall'indice di sviluppo distinto per genere. I livelli di significatività osservati indicano che non vi è differenza tra le due distribuzioni, per nessun indice di sviluppo.

Per quanto riguarda la nazionalità dei soggetti, invece, bisogna tener presente che i soggetti non italiani costituiscono solamente il 3% del campione, ovvero sono 50 su 1468. Dai boxplot in Figura 2.13 si può notare che vi sono molti valori anomali per l'indice di sviluppo motorio associato ai bambini italiani, ma i valori mediani per le due nazionalità sono pressapoco uguali. E' evidente una maggiore differenza tra l'indice di sviluppo comunicativo per i soggetti italiani e quello per gli stranieri, infatti mediamente sembra che i bambini italiani abbiano una capacità comunicativa superiore, anche se di poco. Tuttavia, si osserva che i valori dell'indice di sviluppo comunicativo per i soggetti stranieri sono più "concentrati" e il baffo inferiore del boxplot è molto più corto rispetto a quello relativo all'indice comunicativo per gli italiani. Focalizzandosi sull'indice di sviluppo totale si nota che per gli italiani vi sono molto più outliers in corrispondenza dei valori inferiori.
Come si può notare dai valori dei p-value associati al test di Mann-Whitney, non risultano esserci differenze significative degli indici di sviluppo distinti per nazionalità, fatta eccezione per l'indice comunicativo che si può considerare un "caso-limite" poiché il p-value ottenuto dal test è del 10%.

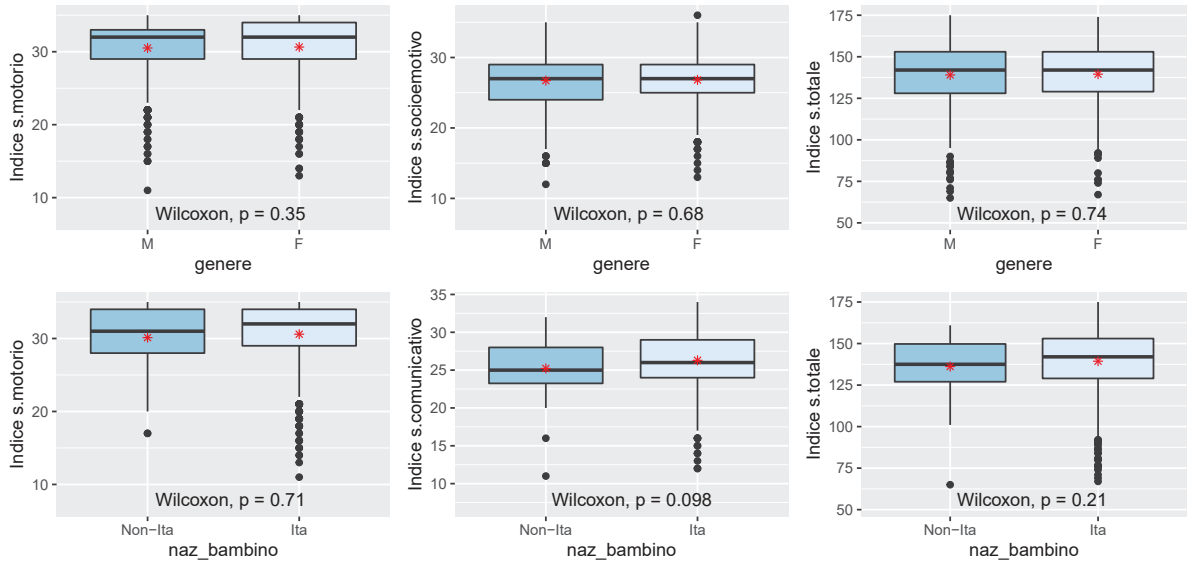

Figura 2.13: Distribuzioni degli indici di sviluppo dei soggetti a seconda del genere e della nazionalità dei soggetti.

A partire dalla Figura 2.14 si analizzano le distribuzioni degli indici di sviluppo motorio, comunicativo e totale a seconda della nazionalità dei genitori. Come per la nazionalità dei soggetti, è importante sottolineare che il numero di bambini aventi i genitori italiani è molto più elevato rispetto a quello di bambini con almeno un genitore straniero. Probabilmente è anche per questo motivo se le distribuzioni degli indici per i soggetti con almeno un genitore non italiano risultano avere differenti asimmetrie, in alcuni casi positive e in altri negative. Si può osservare che i valori mediani di queste distribuzioni risultano essere approssimativamente uguali, infatti i test di Kruskal-Wallis, calcolati per ciascun indice distinto per nazionalità dei genitori, indicano che i gruppi provengono da una stessa popolazione.
La distribuzione dell'indice di sviluppo motorio per i soggetti con genitori italiani presenta una coda sinistra molto più lunga di quella destra, mentre la distribuzione degli altri indici per questo gruppo risulta essere meno asimmetrica ma vi sono sempre outliers in corrispondenza di valori più piccoli.
Per gli altri indici si vedano i boxplot delle distribuzioni distinte per la nazionalità dei genitori in Appendice.

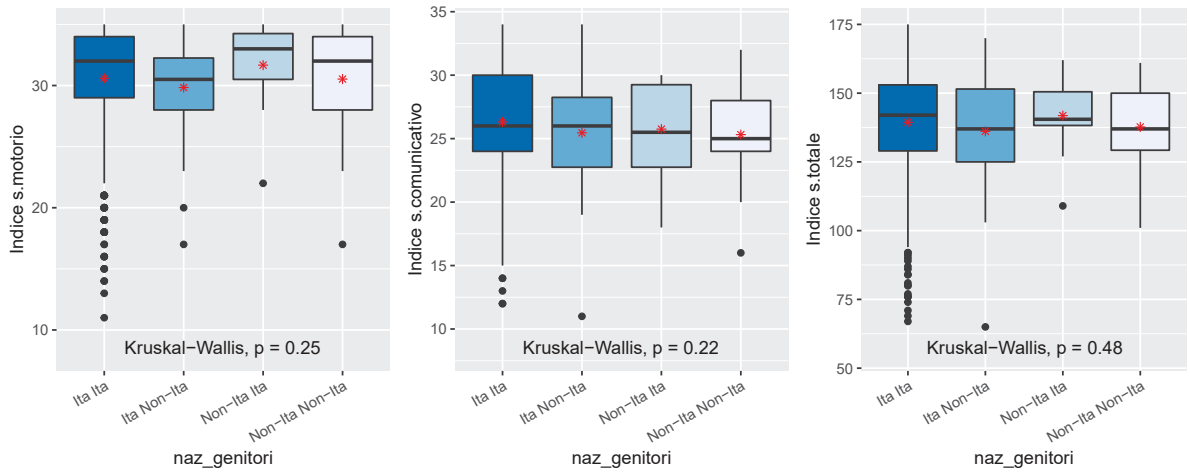

Figura 2.14: Distribuzioni degli indici di sviluppo dei soggetti a seconda della nazionalità dei genitori.

Si analizza ora la relazione degli indici di sviluppo con l'età e la provenienza geografica dei soggetti nel dataset. In Figura 2.15 vengono riportati i boxplot relativi agli indici di sviluppo distinti per Nord, Centro e Sud Italia. Si può osservare che mediamente tutti gli indici hanno valori più bassi e una deviazione standard inferiore nel Centro Italia, mentre tra Nord e Sud la differenza tra i valori mediani delle distribuzioni degli indici sembra meno evidente. Come si può osservare dai p-value calcolati a partire dai test di Kruskal-Wallis per ciascun indice di sviluppo, si ha che i gruppi sono significativamente diversi tra loro, a tutti i livelli α usuali (p-value<0,01). Dall'analisi post-hoc si ottiene che gli indici di sviluppo motorio e adattivo si distiguono nel Centro Italia rispetto alle altre due ripartizioni geografiche, ma non risultano essere significativamente differenti

tra Nord e Sud e Isole; gli altri indici di sviluppo invece, compreso l'indice totale, sono significativamente diversi in ciascuna delle zone geografiche di appartenenza dei soggetti.
In generale le distribuzioni degli indici di sviluppo risultano essere asimmetriche, con valore mediano che si discosta rispetto alla media del gruppo nella maggior parte dei casi, e presentano diversi valori anomali in particolar modo in corrispondenza dei valori più piccoli.

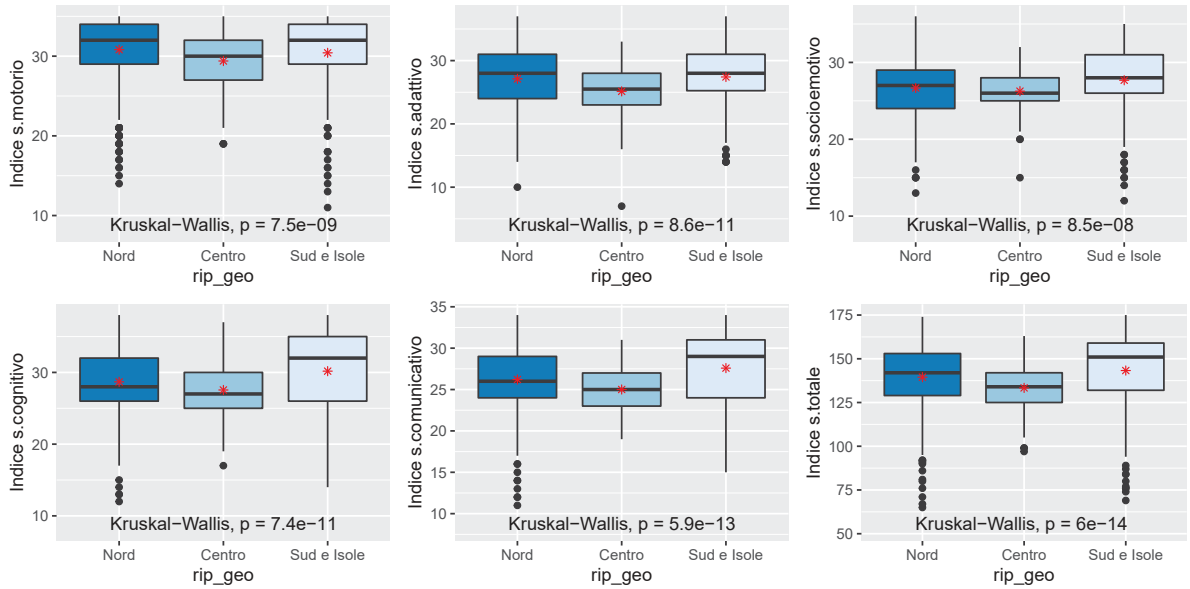

Figura 2.15: Distribuzioni degli indici di sviluppo dei soggetti a seconda della ripartizione geografica.

Dai grafici di dispersione nella Figura 2.16 si può notare che vi è una relazione approssimativamente lineare positiva tra gli indici di sviluppo e l'età dei soggetti appartenenti allo studio, con un'inclinazione più accentuata fino a circa i 50 mesi di età. Questo è ragionevole se si pensa che nei primi anni di età la crescita avviene molto più velocemente.
Si riportano in Appendice i grafici di dispersione degli indici di sviluppo adattivo, socio-emotivo e comunicativo, che mostrano lo stesso andamento lineare appena descritto.
Dal momento che le distribuzioni degli indici di sviluppo non sono risultate essere normali (risultato ottenuto dal test di Shapiro-Wilk condotto nel paragrafo precedente) e nemmeno simmetriche, si è potuta quantificare la relazione esistente tra queste variabili attraverso l'indice di correlazione di Spearman. Esso risulta essere un valore compreso tra 0,6 e 0,83, con un livello di significatività osservato inferiore all'1%, pertanto la correlazione tra i diversi indici e l'età è significativa.

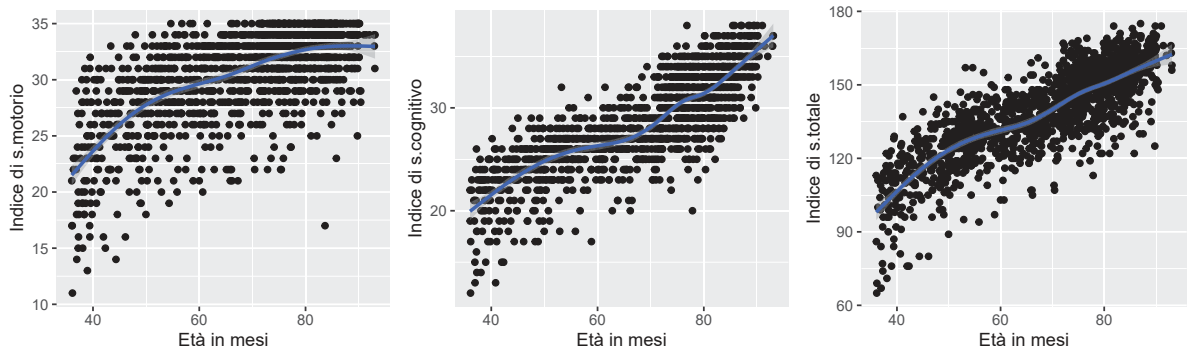

Figura 2.16: Distribuzioni degli indici di sviluppo motorio, cognitivo e totale dei soggetti a seconda dell'età dei soggetti nello studio.

Di seguito si possono osservare i boxplot delle distribuzioni di alcuni indici di sviluppo distinte per lo stato civile dei genitori dei soggetti nel campione.
Dalla Figura 2.17 si osserva che i valori mediani di ciascun indice di sviluppo, a seconda dello stato civile dei genitori, non si discostano molto tra loro. Le distribuzioni degli indici di sviluppo motorio e socioemotivo per le madri separate/divorziate presentano un'asimmetria positiva più rilevante, mentre la distribuzione dell'indice comunicativo per padri celibi mostra un'asimmetria negativa. In generale le distribuzioni degli indici sembrano essere quasi tutte asimmetriche, infatti media e mediana spesso sono piuttosto diverse tra loro. Come indicato in precedenza, si devono tener presente le diverse frequenze per ciascun livello delle variabili relative allo stato civile e la costante presenza di outliers in corrispondenza dei valori inferiori.
Nei grafici sono riportati i p-value associati al test di Kruskal-Wallis calcolato per ciascun indice di sviluppo a seconda dello stato civile della madre e del padre. Essi non risultano essere significativi, con un valore superiore anche al 10%, pertanto gli indici di sviluppo non si differiscono per lo stato civile dei genitori dei soggetti nello studio.
Per gli altri indici si vedano i boxplot delle distribuzioni distinte per lo stato civile dei genitori in Appendice.

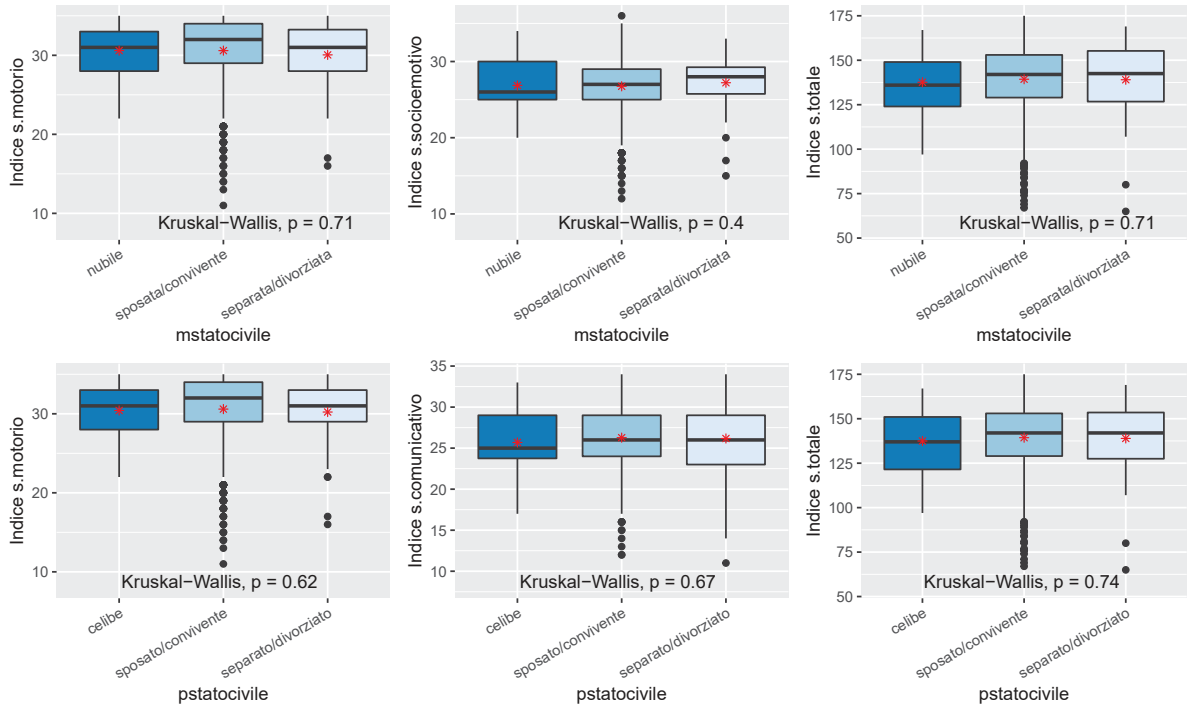

Figura 2.17: Distribuzioni degli indici di sviluppo a seconda dello stato civile dei genitori dei soggetti nello studio.

In Figura 2.18 sono stati riportate le distribuzioni di alcuni indici di sviluppo distinte per titolo massimo conseguito e occupazione lavorativa svolta dai genitori dei soggetti in esame. Gli indici mancanti sono invece riportati in Appendice.
Soffermando l'attenzione sul titolo di studio, si può notare che le distribuzioni dell'indice di sviluppo socioemotivo, distinte per i diversi gruppi, sembrano molto simili tra loro, con valori mediani quasi coincidenti. Il test di Kruskal-Wallis, infatti, fornisce un p-value non significativo al 5% e, pertanto, indica che l'indice socioemotivo non si distingue per titolo di studio dei genitori dei soggetti nello studio.
Per quanto riguarda gli indici di sviluppo cognitivo e totale, invece, si ottiene che i gruppi individuati dal titolo di studio sono significativamente diversi tra loro, con un p-value inferiore all'1%.
Dall'analisi post-hoc effettuata per quegli indici di sviluppo che sono risultati essere differenti a seconda del titolo di studio dei genitori, ovvero tutti tranne l'indice socioemotivo, si è riscontrato che gli indici motorio e adattivo non differiscono tra basso e medio titolo di studio, ma solo tra basso-alto e tra medio-alto livello di istruzione. L'indice comunicativo, invece, non differisce tra alto e medio titolo di studio. Gli indici di sviluppo cognitivo e totale differiscono per tutti i tre i titoli di studio.

I boxplot di ciascun indice di sviluppo a seconda dell'occupazione lavorativa dei genitori dei soggetti, riportati in Figura 2.18, risultano essere approssimativamente uguali tra loro. I soggetti che hanno le madri casalinghe sembrano avere un indice di sviluppo motorio e

cognitivo mediamente più elevato rispetto a coloro che hanno madri full-time, part-time o disoccupate/studentesse. Tale risultato è confermato dal test di Kruskal-Wallis effettuato per questi due indici distinti per occupazione lavorativa della madre, poiché fornisce un p-value inferiore al 5%. Tuttavia, l'analisi post-hoc non parametrica fornisce un p-value significativo solamente per il confronto tra l'indice motorio per i soggetti con madri casalinghe e madri full-time, rifiutando quindi l'ipotesi di uguaglianza tra le mediane di questi due gruppi.
Per gli indici di sviluppo adattivo, socioemotivo, comunicativo e totale, distinti per occupazione della madre dei soggetti, si ottiene che i gruppi provengono dalla stessa popolazione, infatti i p-value associati ai test di Kruskal-Wallis risultano essere non significativi a un livello di significatività del 5%.
Gli indici di sviluppo, distinti per lavoro svolto dal padre dei bambini dello studio, non risultano essere significativamente diversi tra i gruppi, con livelli di significatività osservati tutti più elevati del 10%.
In generale, si osserva che le distribuzioni presentano un'asimmetria più o meno forte, in particolare per quei gruppi in cui si ha un numero molto contenuto di osservazioni si osserva che media e mediana si diversificano maggiormente. Un esempio è il gruppo "part-time" in riferimento all'occupazione del padre, che comprende solamente il 4% del campione.

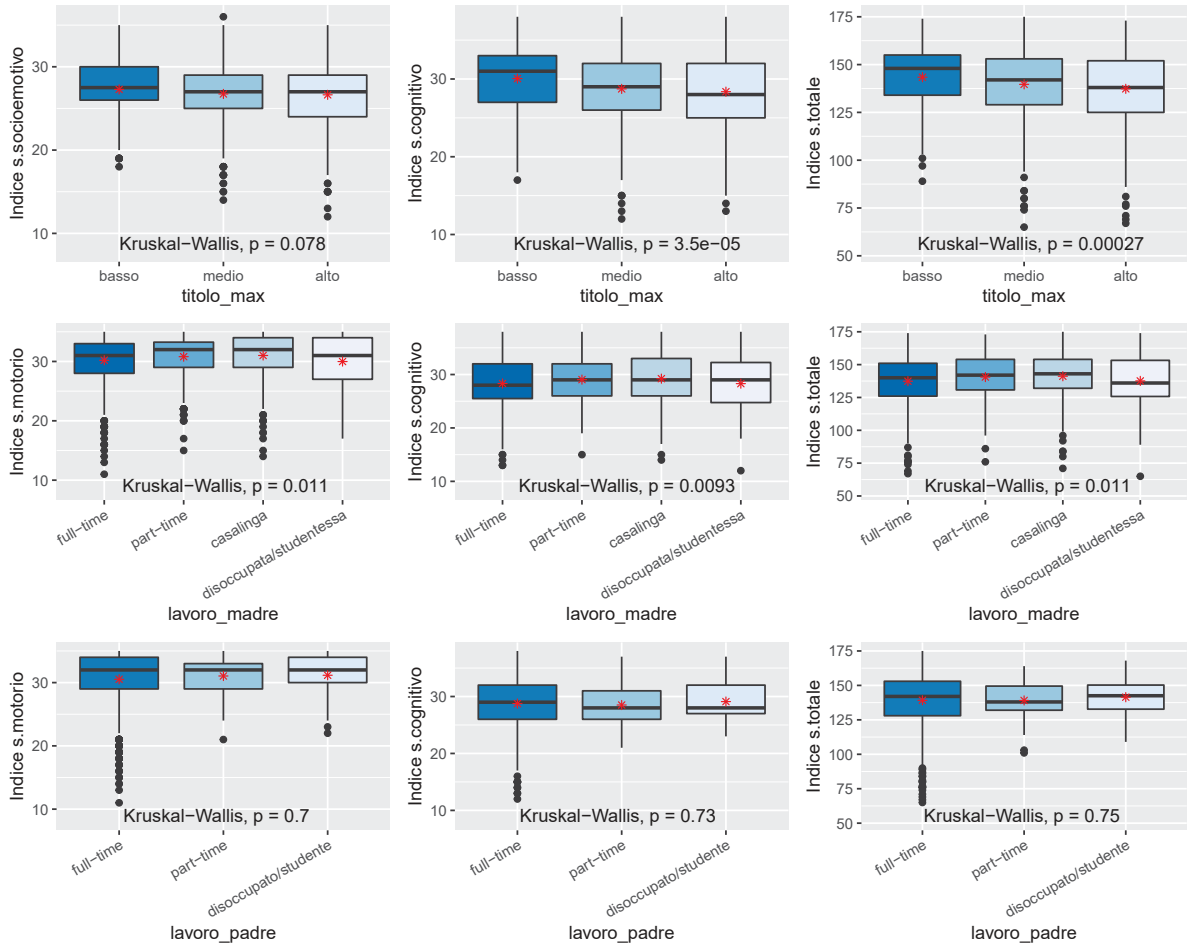

Figura 2.18: Distribuzioni degli indici di sviluppo a seconda del titolo massimo conseguito e dell'occupazione lavorativa dei genitori dei soggetti nello studio.

Dalla Figura 2.19 si possono osservare i boxplot relativi alle distribuzioni degli indici di sviluppo distinte per il numero di figli (oltre al soggetto) e per la presenza di un fratello maggiore dei bambini nello studio. Per quanto riguarda la variabile categoriale *numero di figli*, da una prima analisi grafica si nota che i valori mediani sono leggermente più elevati all'aumentare del numero di figli. Una differenza più marcata è evidente tra la mediana dell'indice motorio per coloro che non hanno altri figli oltre al soggetto, e quella per coloro che hanno invece un altro figlio. Nei grafici sottostanti sono riportati anche i valori dei livelli di significatività osservati per ciascun test effettuato sugli indici di sviluppo a seconda del numero di figli. Essi indicano che vi è una differenza in mediana per ciascun degli indici, distinti per la variabile numero di figli.
E' stata effettuata, inoltre, un'analisi post-hoc per capire quali gruppi si differenziano in distribuzione, ottenendo che, per tutti gli indici di sviluppo, vi è una differenza significativa tra coloro che non hanno altri figli e quelli che hanno almeno un altro figlio oltre al soggetto

in questione, mentre gli indici non sono significativamente diversi per i soggetti che hanno un/una fratello/sorella o più fratelli.
Riassumendo, si osserva che gli indici di sviluppo si differenziano principalmente tra coloro che hanno un solo figlio e coloro che hanno almeno due figli oltre al soggetto considerato nello studio.
Dai boxplot distinti per la presenza di un fratello minore per i soggetti, si può osservare che vi è una differenza significativa tra i gruppi individuati per ciascun indice di sviluppo, con p-value inferiori all'1%.

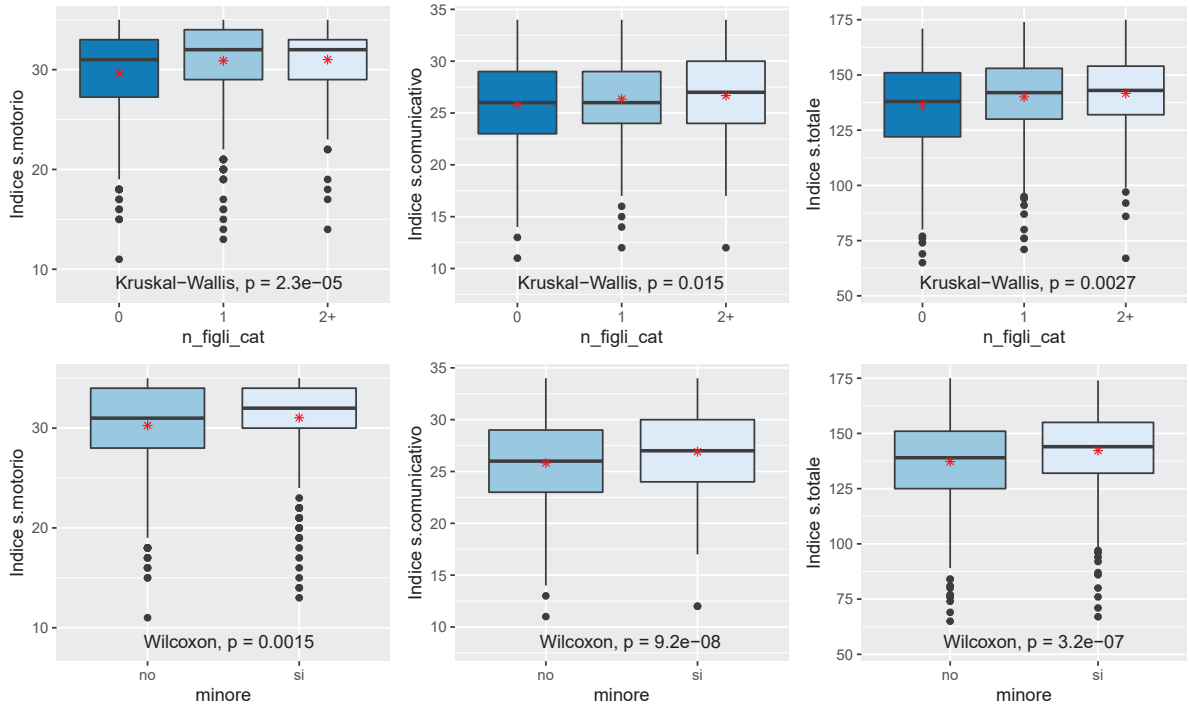

Figura 2.19: Distribuzioni degli indici di sviluppo a seconda del numero di figli, compresi i soggetti nello studio, e a seconda che i soggetti abbiano almeno un fratello minore.

Successivamente si è studiata la relazione tra i diversi indici di sviluppo e le variabili relative alla raccolta dei dati. Dalla Figura 2.20 si può osservare che la mediana dell'indice di sviluppo motorio per il gruppo "intervista" sembra essere inferiore rispetto a quella per il gruppo "questionario", mentre le mediane degli indici cognitivo e totale non sembrano differirsi tra i due gruppi. Tale risultato è confermato dai valori dei p-value calcolati dai test di Mann-Whitney, riportati nei grafici sottostanti.
In generale le distribuzioni sono asimmetriche, in particolare i boxplot dell'indice di sviluppo motorio, distinti per tipo di raccolta, hanno baffi superiori molto più corti di quelli inferiori e presentano un numero considerevole di outliers.

Focalizzando l'attenzione sui grafici delle distribuzioni degli indici di sviluppo a seconda della variabile relativa al luogo di raccolta, si nota che vi è una differenza significativa tra le mediane dei punteggi degli indici ottenuti dai soggetti, infatti i livelli di significatività osservati confermano questa conclusione.
Si osserva che le distribuzioni degli indici di sviluppo per i dati raccolti presso il luogo di lavoro sono molto asimmetriche rispetto agli altri gruppi in questione, tuttavia è necessario tener presente che si hanno solamente 7 osservazioni su 1468. Probabilmente sarebbe più opportuno non considerare la sede di lavoro tra le modalità della variabile *luogo* di raccolta dei dati.

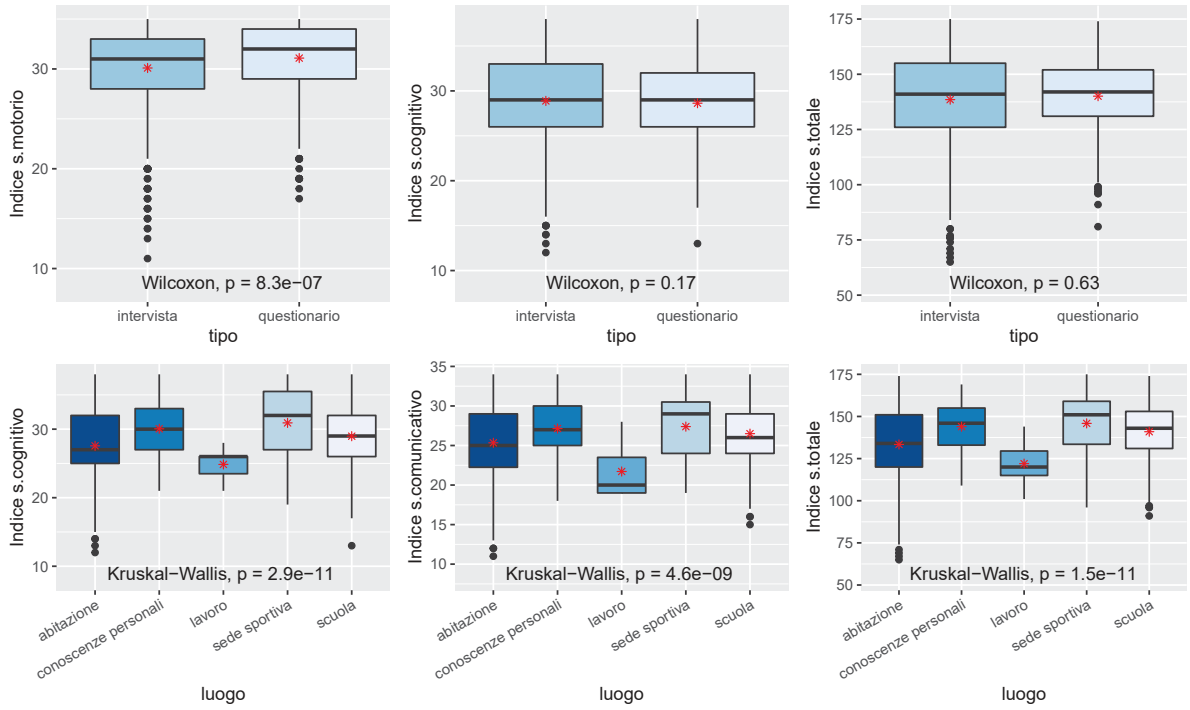

Figura 2.20: Distribuzioni degli indici di sviluppo a seconda del tipo e del luogo di raccolta dei dati.

In Figura 2.21 si può osservare come le distribuzioni degli indici di sviluppo possono differenziarsi a seconda della persona che raccoglie i dati relativi ai soggetti dello studio. Infatti, la valutazione dell'aspetto cognitivo o comunicativo può essere talvolta soggettiva, ovvero non ci sono delle regole standard per poter dare un punteggio di sviluppo unico a un bambino. Dai boxplot si osserva la presenza di una variabilità individuale: alcuni intervistatori danno un punteggio molto limitato, come accade per i rilevatori 12, 14 e 21, altri invece valutano i soggetti in modo molto diverso l'uno dall'altro, come accade, ad esempio, per i rilevatori 9 e 19. Si osserva, inoltre, che alcuni rilevatori associano indici più elevati ai soggetti, mentre altri valutano i bambini con indici inferiori.
Tale situazione è dovuta al *bias* di rilevazione, molto ricorrente negli studi epidemiologici.

Il *bias* o distorsione è una modifica, intenzionale o non intenzionale, del disegno e/o della conduzione di uno studio che comporta un'errata valutazione dei dati. [9]
Per quanto riguarda lo studio considerato in questa tesi, il *bias* avviene nell'intervista, ovvero vi è una differenza sistematica nel sollecitare, registrare o interpretare le informazioni dai partecipanti allo studio. [10]
Il test di Kruskal-Wallis effettuato per gli indici di sviluppo cognitivo e comunicativo a seconda del rilevatore rifiuta l'ipotesi di omogeneità fra i gruppi a tutti i livelli α usuali, dunque conferma quanto si vede dai boxplot in Figura 2.21.

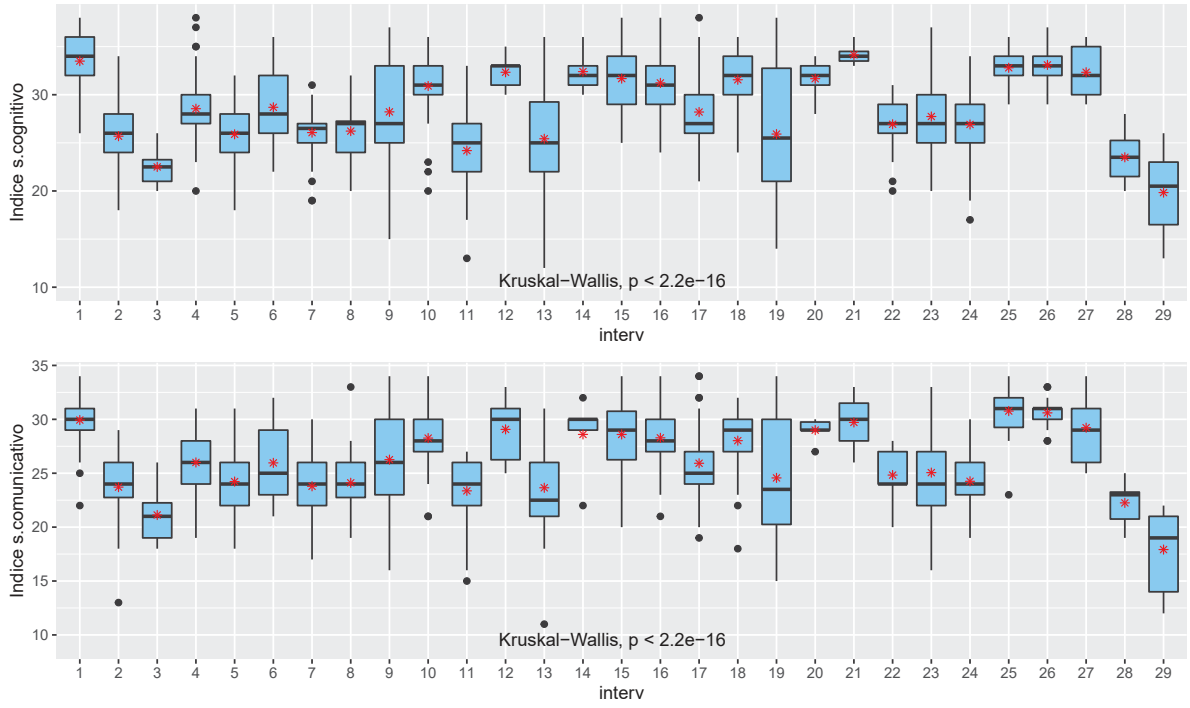

Figura 2.21: Distribuzioni degli indici di sviluppo cognitivo e comunicativo a seconda della persona che ha rilevato i dati.

Dall'analisi svolta fino a questo momento si è potuto osservare come e quanto gli indici di sviluppo dei soggetti si associano alle altre variabili presenti nel dataset. In particolare, tra le caratteristiche demografiche dei soggetti, la ripartizione geografica di appartenenza e l'età influenzano molto gli indici di sviluppo.
Inoltre, in merito alle caratteristiche relative ai genitori dei bambini, quelle che si associano significativamente alle variabili di interesse sono il titolo massimo conseguito, l'occupazione lavorativa della madre (anche se in misura molto inferiore) e il numero di figli.
Infine, anche le variabili relative al luogo di raccolta dei dati e all'intervistatore sembrano influenzare i punteggi degli indici dei soggetti in esame.

3 Il modello fattoriale: indici di sviluppo e PM_{10}

Come si è già visto nel capitolo precedente, le variabili di interesse sono gli indici di sviluppo associati ai soggetti dello studio. Per poter facilitare le analisi e avere un quadro generale più comprensibile si può considerare l'indice di sviluppo totale come unica variabile risposta, poiché esso deriva dalla somma degli indici motorio, adattivo, socioemotivo, cognitivo e comunicativo. Tuttavia, tale indice è una misura "grezza" e non è sempre una buona sintesi delle informazioni derivanti dai singoli punteggi.
Per poter svolgere le analisi in modo più accurato, in questo capitolo si cerca di costruire un indice opportuno che riassuma l'informazione proveniente da ciascuno dei 5 indici a disposizione, attraverso un'analisi dei fattori. In questo modo si otterrà un'unica variabile, ovvero un fattore comune, derivante dalla combinazione lineare dei singoli indici di sviluppo con coefficienti detti *loadings*.

3.1 L'analisi fattoriale

L'analisi fattoriale (FA) è un modello sulla matrice di covarianza dei dati, basato sull'ipotesi della normalità multivariata del vettore delle variabili, e la funzione di verosimiglianza che ne consegue viene utilizzata per la stima dei parametri. Nella FA si conosce a priori il numero di variabili (fattori) che si vogliono ottenere.
Per un generico vettore aleatorio $\underset{(p\times 1)}{X}$, si assume: [11]

- $X \sim f_p(\mu, \Sigma)$, dove $f_p()$ è una certa distribuzione *p-variata*, con
- $\mu = E(X)$ il vettore $(p \times 1)$ delle medie,
- $\Sigma = Cov(X)$ la matrice $(p \times p)$ di covarianza di X

Il modello fattoriale assume che il vettore X dipenda linearmente da (possibilmente) poche variabili latenti $F_1, F_2, ..., F_m$ detti *fattori comuni* e altre p variabili $\epsilon_1, \epsilon_2, ..., \epsilon_p$ detti *errori* o *fattori specifici*. In particolare, il modello della FA è

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + ... + l_{2m}F_m + \epsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + \epsilon_p \end{aligned}$$

In forma compatta il modello fattoriale diventa

$$\underset{(p\times 1)}{X} - \mu = \underset{(p\times m)}{L}\underset{(m\times 1)}{F} + \underset{(p\times 1)}{\epsilon} \tag{1}$$

dove L è la matrice dei *loadings*, quindi l_{ij} rappresenta il peso della i-esima variabile nel fattore j, $i \leq p$, $j \leq m$. Poiché tale modello così presentato ha troppi parametri ignoti, è opportuno fare delle ipotesi più restrittive. Si considera, dunque, un modello a fattori ortogonali, il quale assume:

- $E(F) = \underset{(m\times 1)}{0}$, $Cov(F) = E(FF') = \underset{(m\times m)}{I}$
- $E(\epsilon) = \underset{(p\times 1)}{0}$, $Cov(\epsilon) = E(\epsilon\epsilon') = \underset{(p\times p)}{\Psi} = diag(\psi_1, \psi_2, ..., \psi_p)$
- $Cov(F, \epsilon) = E(F\epsilon') = \underset{(m\times p)}{0}$

Tale modello implica una struttura specifica per Σ, la matrice di covarianza di X. Infatti, ne consegue che

$$\Sigma = Cov(X) = E[(X-\mu)(X-\mu)'] = LL' + \Psi \quad e \quad Cov(X, F) = L \tag{2}$$

In sostanza, la (2) indica che:

$$V(X_i) = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2 + \psi_i$$

$$Cov(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + ... + l_{im}l_{km} e Cov(X_i, F_j) = l_{ij}$$

Nella FA si è soliti porre $h_i^2 = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2$, e h_i^2 è chiamata i-esima comunalità poiché costituisce la porzione della varianza totale di X_i attribuita agli m fattori comuni. La ψ_i è detta varianza specifica (o specificità), pertanto

$$\sigma_{ii} = h_i^2 + \psi_i = comunalità + varianza\, specifica, \quad i = 1, 2, ..., p$$

Ora, considerando il vettore $\underset{(5\times 1)}{X} = (X_1, X_2, X_3, X_4, X_5)$ degli indici di sviluppo, si stima un modello fattoriale a partire da S, ovvero la stima campionaria di Σ, e attraverso la stima di massima verosimiglianza dei parametri L e Ψ.
In particolare, si vuole ottenere un solo fattore $F = F_1$ e quindi si avrà un vettore $L = (l_1, ..., l_5)$ dei loadings, il quale è riportato nella Tabella 3.1 sottostante. La varianza totale del vettore X, dato dai 5 indici di sviluppo, attribuita al fattore F_1 è data dalla somma dei quadrati dei loadings ed è pari a 3,657, ovvero la porzione di varianza totale attribuita al primo fattore è pari al 73% della varianza totale.
Dall'analisi fattoriale svolta con il metodo della massima verosimiglianza si ottengono, inoltre, le specificità stimate (indicate con *uniqueness*), le quali vengono riportate nella Tabella 3.2.

Uniqueness				
motorio	adattivo	socioemotivo	cognitivo	comunicativo
0.378	0.311	0.329	0.165	0.159

Tabella 3.1: Stime di massima verosimiglianza delle varianze specifiche, ottenute attraverso l'analisi fattoriale.

	Loadings				
	motorio	adattivo	socioemotivo	cognitivo	comunicativo
Factor1	0.788	0.83	0.819	0.914	0.917

Tabella 3.2: Stime di massima verosimiglianza dei loadings, ottenute attraverso l'analisi fattoriale.

Le correlazioni tra gli indici di sviluppo motorio, adattivo, socioemotivo, cognitivo e comunicativo risultano essere significative, infatti hanno valori che variano tra 0,671 e 0,856. Dalla Figura 3.1 si può osservare la matrice dei grafici a dispersione di ciascuna coppia di indici di sviluppo a sinistra, mentre le correlazioni di Pearson si possono visualizzare a destra.

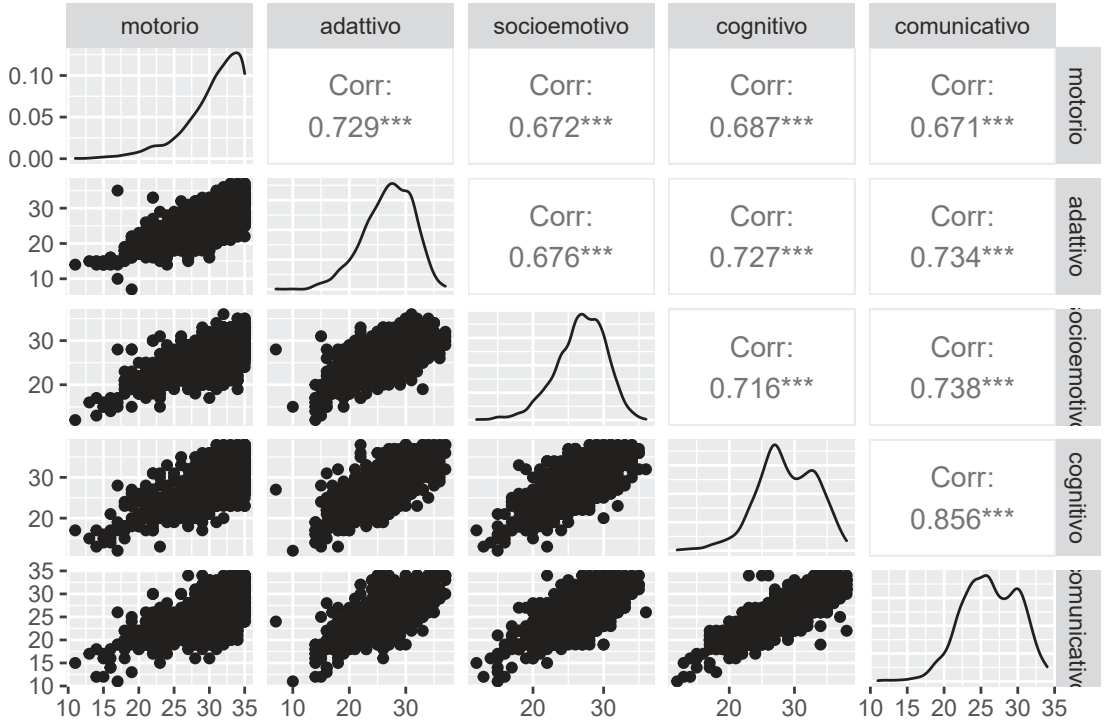

Figura 3.1: Matrice dei grafici a dispersione di ciascuna coppia degli indici di sviluppo, con relative correlazioni di Pearson.

Per poter costruire un'unica variabile che riassuma le informazioni fornite da ciascuno dei 5 indici di sviluppo originali, si considera una combinazione lineare di questi indici standardizzati dove i pesi sono i loadings ottenuti dall'analisi fattoriale. Per ottenere tale combinazione, che verrà indicata *score*, è necessario ricorrere ai *regression score* (punteggi fattoriali) ottenuti tramite l'analisi dei fattori. In generale, gli *score* di regressione sono così definiti:

$$\underset{(n\times m)}{\hat{F}} = \underset{(n\times p)}{Z}\underset{(p\times m)}{B}$$

dove F è la matrice dei punteggi fattoriali, di dimensioni $(n \times m)$ in cui n è il numero di unità statistiche e m è il numero di fattori (nel nostro caso m=1); Z è la matrice delle variabili di partenza standardizzate, con p numero di variabili; B è la matrice dei pesi fattoriali, ovvero dei *loadings*, attribuiti a ciascun fattore comune.[12]
Ciascun punteggio fattoriale costruito in questo modo possiede alcune proprietà importanti: esso ha media nulla e varianza pari alla correlazione multipla al quadrato (SMC) tra ciascuna variabile e i fattori, ovvero le comunalità. Considerando un unico fattore, lo score è dunque un vettore $(n \times 1)$, assume valori compresi fra circa -4 e 2 con valore medio pari a 0 e vi sono diversi valori anomali in corrispondenza dei valori negativi inferiori. Tali valori anomali dello score individuano quei soggetti per i quali si hanno indici di sviluppo particolarmente sotto la media.
I *regression score*, infatti, assegnano a ciascuna unità statistica un punteggio che indica dove si colloca quel soggetto in relazione agli indici di sviluppo. Pertanto se un bambino dello studio ha indici di sviluppo molto bassi, egli avrà un punteggio più basso, mentre chi ha ottenuto degli indici di sviluppo intorno alla media avrà uno score all'incirca nullo.

3.2 Indici di sviluppo e la variabile età

Tenendo in considerazione quanto visto nel Capitolo 2, non si riportano nuovamente le analisi bivariate per lo score e l'indice totale. Tuttavia, è fondamentale notare come gli indici di sviluppo differiscono al crescere dell'età. Come si è potuto osservare dalle analisi effettuate precedentemente, infatti, vi è una forte associazione tra le variabili di interesse e l'età dei soggetti dello studio.
Nella Tabella 3.3 si possono visualizzare alcune statistiche di sintesi dello score e dell'indice totale distinti per le classi d'età, ovvero sono riportate le mediane di ciascun gruppo con il relativo intervallo interquartile. Le due variabili sono molto diverse fra loro in quanto sono state costruite in modi differenti, ma il comportamento è molto simile. Il linea generale, infatti, si osserva un andamento crescente e lineare dello score e dell'indice totale all'aumentare dell'età. Tale risultato è evidente anche dai valori significativi dei p-value (inferiori all'1%) associati ai test di Kruskal-Wallis effettuati per ciascuna delle due variabili.
A differenza di quanto fatto nelle analisi bivariate, qui si è considerata la variabile età sud-

divisa in 5 classi anzichè 10. In questo modo, infatti, è più facile dare un'interpretazione ai dati.

	Classi di età					
Variabile	[36,48], N = 169	(48,60], N = 251	(60,72], N = 287	(72,84], N = 564	(84,96], N = 197	p-value
totale	112 (104, 120)	128 (119, 134)	136 (130, 143)	150 (142, 156)	158 (152, 162)	<0.001
score	-1.38 (-1.86, -1.06)	-0.65 (-0.99, -0.33)	-0.24 (-0.57, 0.10)	0.59 (0.14, 0.97)	1.12 (0.79, 1.29)	<0.001

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

Tabella 3.3: Statistiche riassuntive per l'indice di sviluppo totale e per lo score a seconda delle classi d'età dei soggetti.

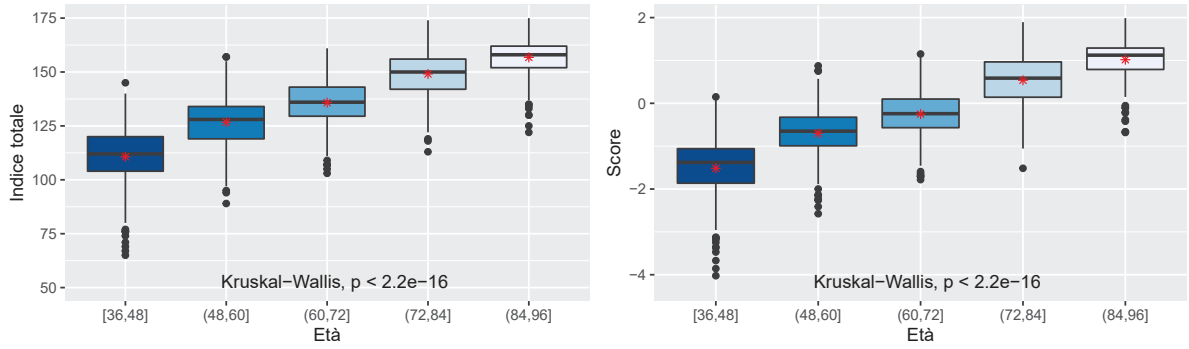

Figura 3.2: Grafico dell'indice di sviluppo totale e dello score a seconda delle classi d'età dei soggetti.

3.3 Indici di sviluppo e PM_{10}

Obiettivo principale di questa tesi è capire se esiste una relazione fra gli indici di sviluppo dei soggetti in esame e le concentrazioni di PM_{10} durante la gravidanza. In questo paragrafo, in un primo momento si analizza la correlazione marginale, poi, alla luce delle considerazioni fatte per l'età dei soggetti, si studia l'associazione tra le due variabili di interesse (indice totale e score) e la concentrazione di PM_{10} a seconda delle classi di età.
Come si può osservare dai grafici a dispersione in Figura 3.3, non vi è un'associazione marginale significativa tra lo score e il livello di PM_{10} in ciascun mese di gravidanza. Tale risultato è confermato dai coefficienti di Spearman calcolati per ciascun mese, i quali risultano essere valori compresi tra 0,06 e 0,16.
Considerando l'indice di sviluppo totale, dalla Figura 3.4 si osserva che, anche in questo caso, non vi è una correlazione significativa tra questa variabile e la concentrazione di PM_{10} in ciascun mese di gravidanza, con coefficienti di correlazione di Spearman tra 0,06 e 0,16.

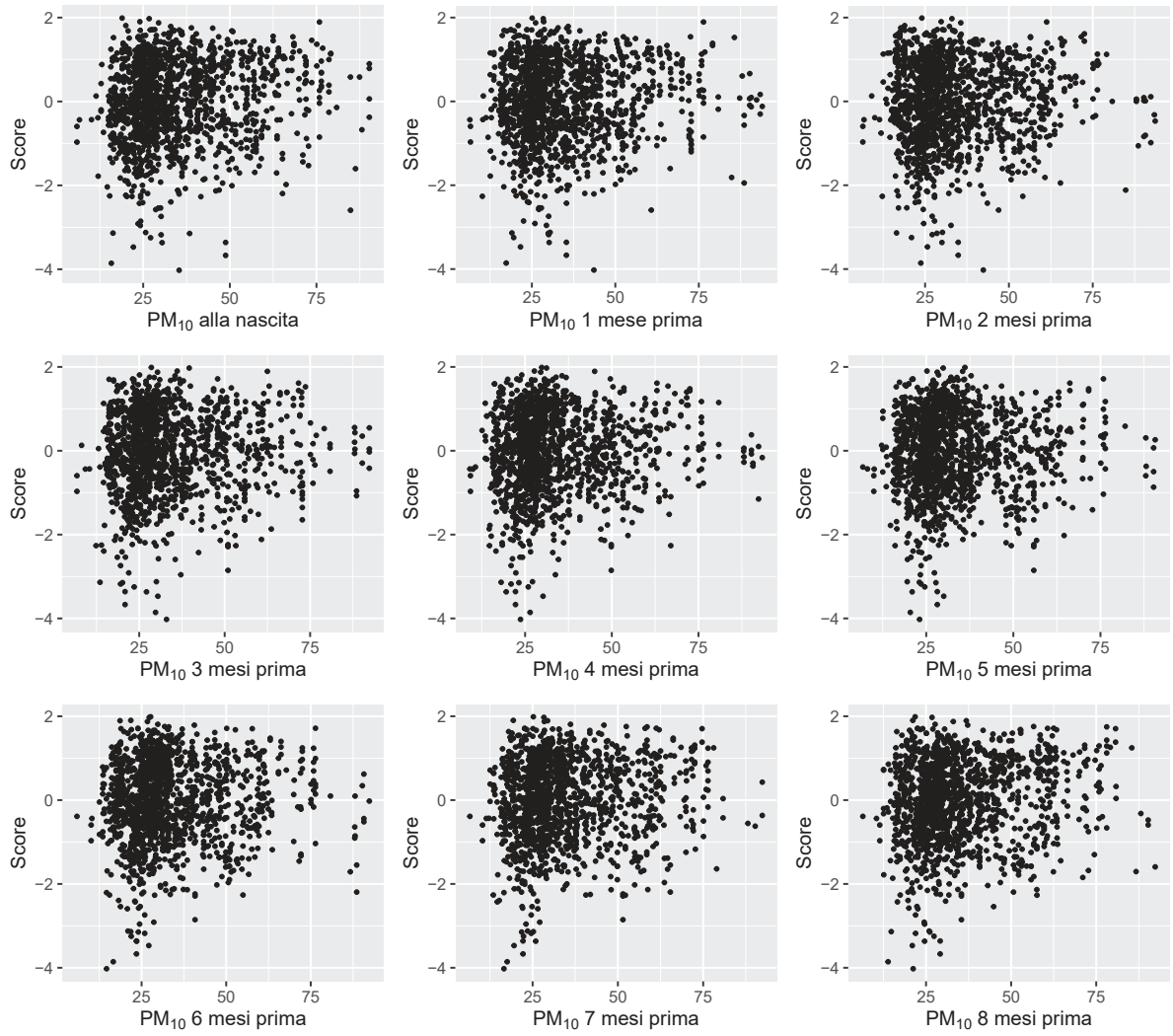

Figura 3.3: Distribuzione dello score a seconda del livello di PM_{10} alla nascita e in ciascun mese precedente alla nascita.

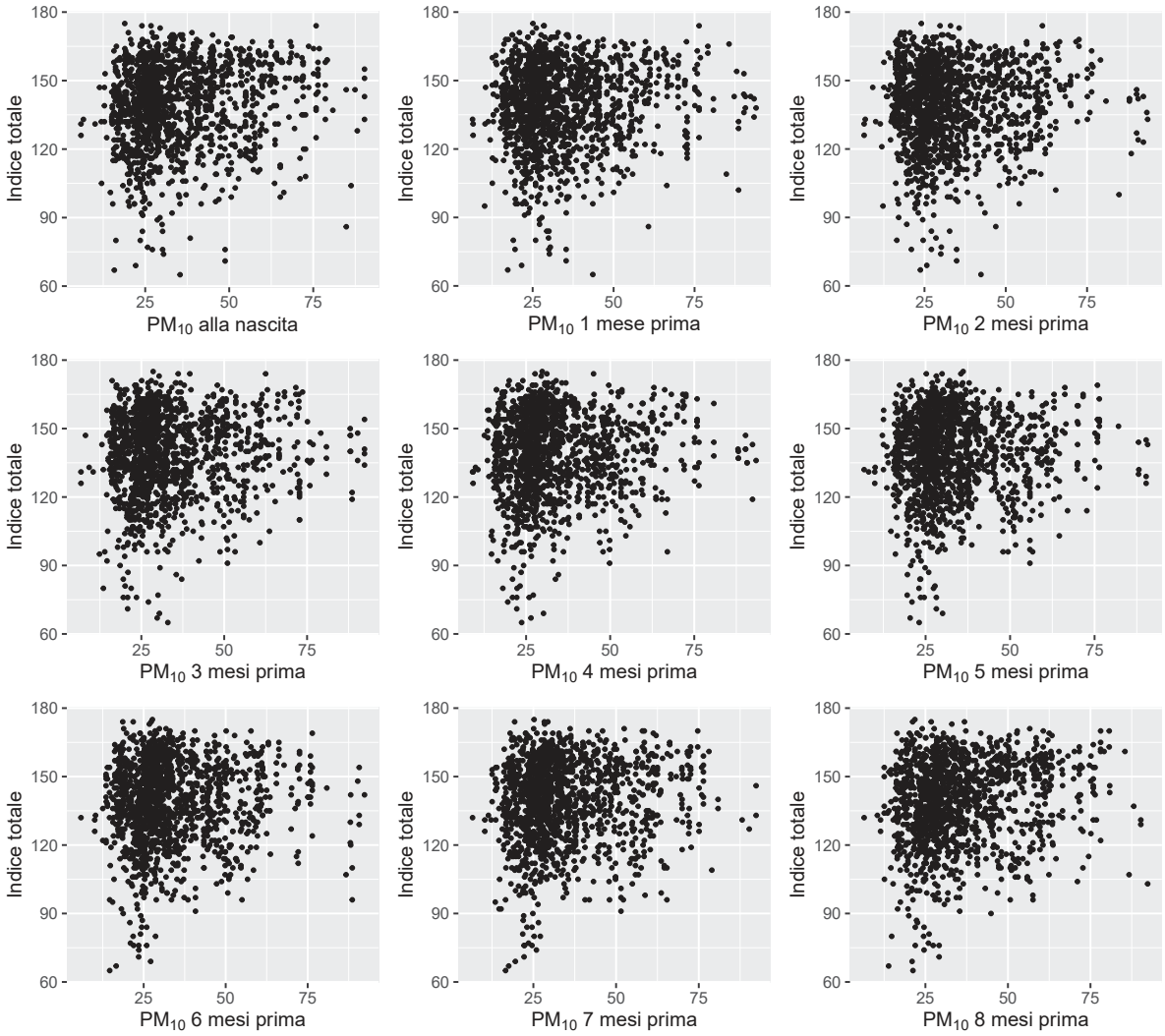

Figura 3.4: Distribuzione dell'indice di sviluppo totale dei soggetti a seconda del livello di PM_{10} alla nascita e in ciascun mese precedente alla nascita.

Si considera ora la relazione dell'indice di sviluppo totale e dello score con le concentrazioni di PM_{10}, per ciascuna classe di età dei soggetti nello studio. Come si può osservare dai grafici a dispersione in Figura 3.5 e 3.6, si può dire che vi è una correlazione quasi nulla tra le variabili di interesse e il livello di PM_{10} in ciascun mese di gravidanza, fissata la classe di età a cui appartengono i soggetti. Tale assenza di associazione (lineare) è dovuta al fatto che l'età risulta essere un confondente nella relazione in esame.
Nonostante non vi sia un'associazione tra le variabili considerate, si può ossevare che, tra i bambini più piccoli, ovvero quelli di età compresa tra i 36 e i 48 mesi, quelli con indice di sviluppo totale e score particolarmente bassi sono stati esposti a concentrazioni di PM_{10} inferiori durante la gravidanza.

In linea generale i punti dei grafici a dispersione non si distribuiscono seguendo un andamento lineare positivo o negativo evidente, ovvero all'aumentare dei livelli di inquinamento non sembra conseguire una modifica dei valori delle variabili di interesse. Pertanto si può affermare che la relazione dell'indice totale e dello score con le concentrazioni di PM_{10}, fissata la classe d'età a cui appartengono i soggetti, è sostanzialmente nulla.

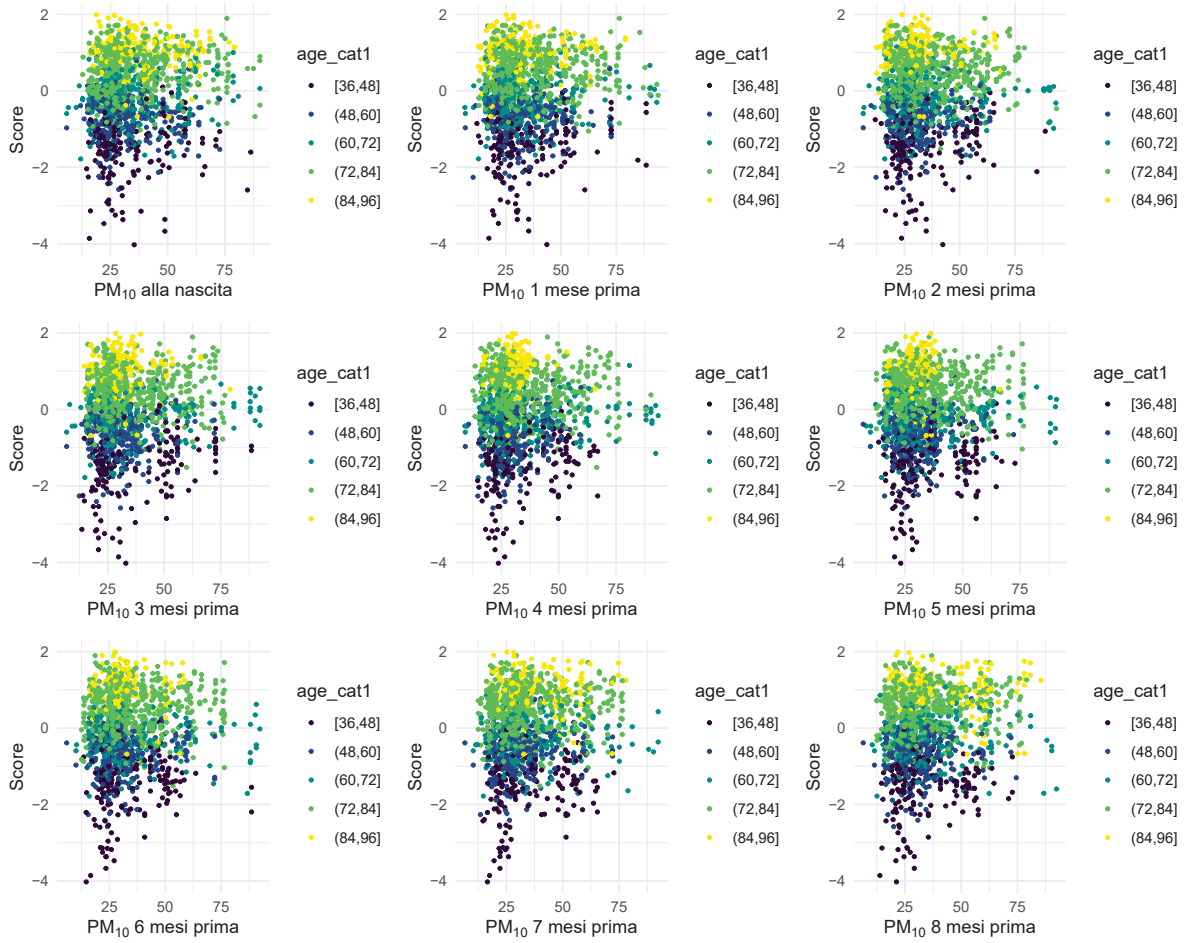

Figura 3.5: Distribuzione dello score dei soggetti a seconda del livello di PM_{10} in ciascun mese di gravidanza e distinto per classi d'età dei soggetti.

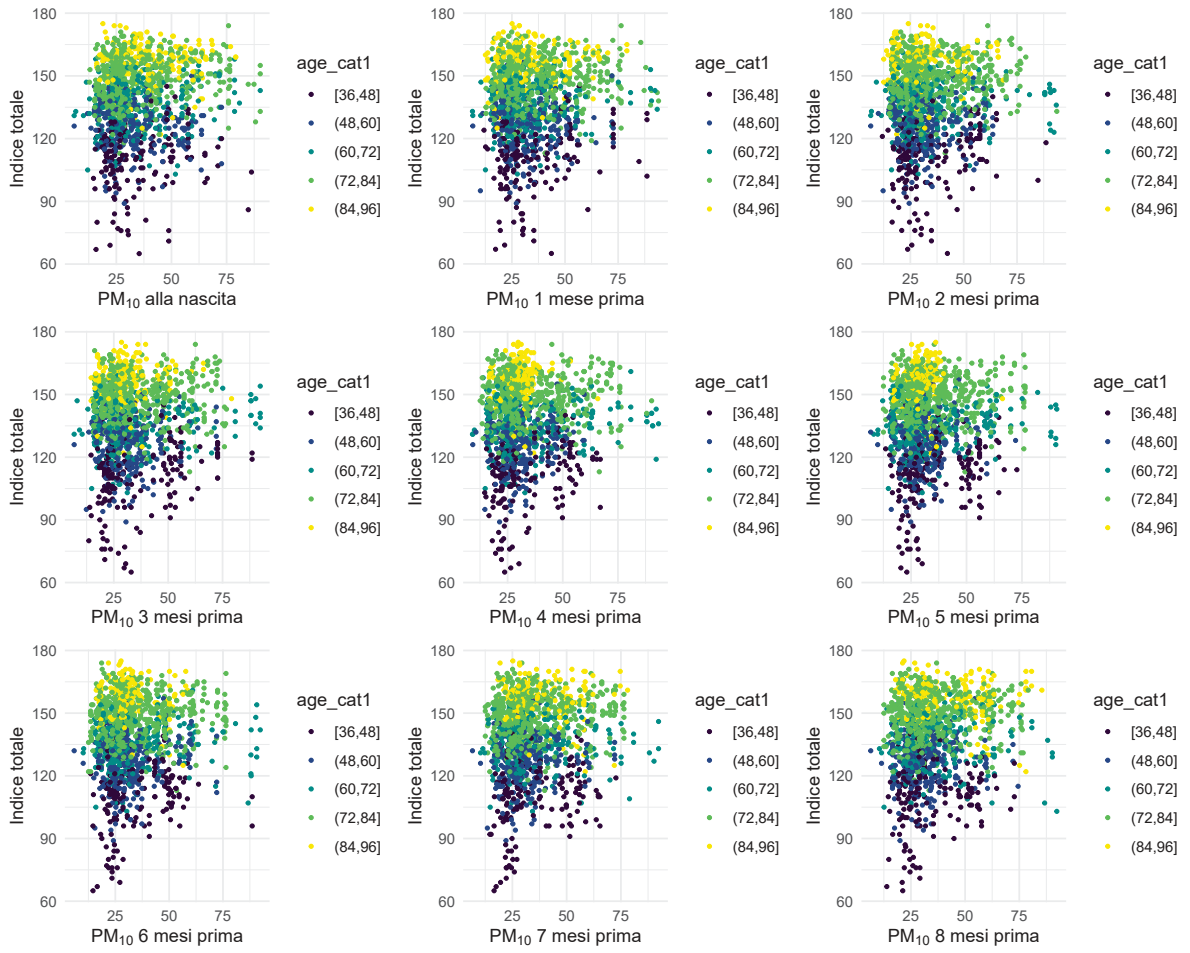

Figura 3.6: Distribuzione dell'indice di sviluppo totale dei soggetti a seconda del livello di PM_{10} in ciascun mese di gravidanza e distinto per classi d'età dei soggetti.

4 L'analisi dei dati funzionali

4.1 Cosa sono i dati funzionali

In alcuni contesti, come nel caso di una covariata relativa all'inquinamento da PM_{10}, per ciascuna unità statistica sono presenti più osservazioni nel tempo della stessa variabile. Nel dataset analizzato, infatti, si hanno a disposizione misure ripetute delle concentrazioni di PM_{10}, ossia per ciascuno dei 9 mesi di gravidanza sono stati raccolti i livelli dell'agente inquinante (valori medi) in riferimento ad ogni soggetto dello studio.
In tali circostanze ogni unità statistica è rappresentata da una serie temporale di valori riferiti a una o più variabili. Se le osservazioni sono sufficientemente frequenti nel tempo, si possono identificare come la realizzazione a tempi finiti di un processo continuo (funzione) e utilizzare successivamente, ad esempio in un modello, questa covariata come una funzione.
Il termine *funzionale* è relativo alla struttura intrinseca del dato, dunque l'osservazione funzionale x_i per l'i-esimo individuo proviene da un processo del tipo

$$x_i(t) = f_i(t) + \epsilon_i, \qquad t = 1, ..., T, \quad i = 1, ..., n$$

in cui ϵ_i è un errore casuale a media nulla. Pertanto, se i dati risultano essere affetti da errore, è possibile attuare un processo di *interpolazione* o di *smoothing*. A tale scopo un approccio flessibile può essere quello basato su *basi di splines*. Nel paragrafo seguente vengono introdotte le principali basi a cui si ricorre solitamente nell'analisi dei dati funzionali.

4.2 Le basi

Una *base* è una combinazione lineare di K funzioni ϕ_k, linearmente indipendenti tra loro, che permette di approssimare una data funzione. In generale le basi di funzioni possono quindi rappresentare l'andamento di dati funzionali, pertanto essi si possono esprimere nel seguente modo:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

dove c_k sono i coefficienti reali e ϕ_k le funzioni di base. I coefficienti vengono determinati utilizzando il criterio dei minimi quadrati.
L'adattamento perfetto ai dati viene raggiunto quando $K = n$, ovvero quando si scelgono i coefficienti ϕ_k in modo tale da ottenere $x(t_j) = y_j$ per ogni j $(j = 1, ..., n)$. In questo caso, però, si ottiene un adattamento interpolatorio che cambia direzione di movimento senza alcun vincolo. La scelta del numero K è fondamentale per capire il grado ottimale di liscezza della funzione base, in modo che l'adattamento non risulti nè troppo *smoothed* (caso limite è la regressione lineare) nè troppo interpolatorio. Dunque K deve essere il minimo possibile ma deve anche approssimare bene i dati, pertanto risulta essere esso stesso un

parametro da stimare a seconda delle caratteristiche dei dati.
Il sistema di funzioni di base maggiormente noto è la base polinomiale

$$1, t, t^2, t^3, ..., t^k, ...$$

Essa rappresenta un buon punto di partenza, ma spesso risulta essere poco efficiente in termini inferenziali rispetto ad altre basi. La base polinomiale non riesce a interpolare localmente in modo preciso se non si utilizza un valore di K molto grande, inoltre tendono ad approssimare bene al centro dell'intervallo e peggio agli estremi, per questo motivo non sono indicate per la previsione o l'estrapolazione.
Altre basi di funzioni particolarmente note in letteratura sono, ad esempio, quelle ricavate dalla serie di Fourier, la basi Spline e le B-Spline.

Base di Fourier

Le basi derivanti dalla serie di Fourier sono utilizzate per le funzioni di tipo periodico e si esprimono nel seguente modo:

$$1, \sin(wt), \cos(wt), \sin(2wt), \cos(2wt), ...$$

con la costante ω in relazione $\omega = 2\pi/T$ con il periodo T^2. Tale sistema di basi è utilizzato soprattutto in presenza di funzioni regolari dove non sono presenti forti sistematicità locali e dove le curve sono tendenzialmente dello stesso ordine in tutta la funzione. Nonostante questo, in certi casi le basi di Fourier vengono utilizzate anche per approssimare funzioni non periodiche e non regolari in virtù della loro versatilità e semplicità di calcolo.

Spline

Le *spline* sono il sistema di basi più comune per approssimare funzioni non periodiche, infatti combinano la facilità computazionale dei polinomi con la capacità di cambiare comportamento localmente. Per tale motivo, le spline sono uno dei sistemi di basi più flessibili, poiché sono in grado di interpolare andamenti locali delle curve sfruttando l'idea di costruire delle funzioni polinomiali a tratti.
Il primo passo per definire una spline è quello di suddividere l'intervallo su cui la funzione da approssimare è definita in un certo numero di sotto-intervalli, separati da punti distinti τ_k detti *nodi*. La base spline è quindi composta di polinomi di grado fissato d, uno per ciascun sotto-intervallo, e tali polinomi vengono uniti in corrispondenza dei nodi. L'unico vincolo da rispettare è che polinomi adiacenti devono avere un opportuno grado di lisciamento nei nodi. Per tale motivo si sceglie una funzione che abbia derivate dal grado 0 al grado $d - 1$ continue in ogni punto τ_k, in modo che non sia percepibile la discontinuità presente in corrispondenza dei nodi. Solitamente con un grado dei polinomi pari a 3 si riesce a garantire la continuità delle prime due derivate e ottenere una curva liscia.

Il numero di nodi può essere scelto arbitrariamente oppure attraverso criteri di selezione specifici come il criterio informativo di Akaike (AIC) o il criterio di informazione bayesiano (BIC), mentre la relazione che regola la scelta del numero di basi è:

$$numero\,di\,basi = grado\,del\,polinomio + numero\,di\,nodi\,interni$$

dove i nodi interni sono tutti i nodi tranne quelli posizionati all'inizio e alla fine del dominio della funzione.
In generale i nodi vengono scelti equispaziati rispetto agli estremi dell'insieme finito che si sta valutando, ma una delle soluzioni più utilizzate è quella di collocare più nodi in quelle regioni in cui la funzione varia maggiormente, e meno nodi dove la funzione ha un comportamento che si discosta leggermente dalla linearità.

B-spline

Nella pratica vengono maggiormente utilizzate le B-spline, le quali non sono altro che un caso particolare delle spline precedentemente descritte. Queste basi presentano una particolare proprietà, detta *compact support property* (Ramsay e Silverman, 2005), ovvero assumono il valore zero ovunque eccetto in un intervallo finito e questo dà un notevole vantaggio dal punto di vista computazionale. Nel caso delle spline cubiche, ad esempio, una B-spline è una spline cubica con supporto nell'intervallo $[\tau_{k-2}, \tau_{k+2}]$.
Considerata una sequenza non decrescente $t = t_j$, la j-esima B-spline normalizzata di ordine k per la sequenza di nodi t si denota come

$$B_{j,k,t}(x) = (t_{j+k} - t_j)[t_j, ..., t_{j+k}](\cdot - x)_+^{k-1} \quad con\, x \in \mathbb{R}$$

dove $(\cdot - x)_+^{k-1}$ indica che la differenza va calcolata mantenendo fisso il valore di x e cambiando di volta in volta il valore t, e $(x)_+$ vale x se esso è positivo, altrimenti vale 0.
Per maggiori informazioni riguardo le B-spline è possibile consultare de Boor (2001).

4.3 Selezione delle variabili confondenti

In questo paragrafo si cerca di individuare quali variabili del dataset risultano avere un'influenza significativa sulla variabile di interesse, ovvero lo score. A tale scopo si costruisce un modello di regressione lineare dove la variabile risposta y_i è lo *score*, indicante lo sviluppo dei soggetti, e le covariate z_i sono le seguenti:

- il genere ($z_1 = 1$ per le femmine);
- la nazionalità ($z_2 = 1$ per gli italiani);
- la nazionalità dei genitori ($z_3 = 1$ per padre italiano e madre non italiana, $z_4 = 1$ per padre non italiano e madre italiana, $z_5 = 1$ per genitori non italiani);

- la ripartizione geografica ($z_6 = 1$ per il Centro, $z_7 = 1$ per il Sud e Isole);
- l'età (z_8, variabile numerica);
- lo stato civile della madre ($z_9 = 1$ per madre sposata/convivente, $z_{10} = 1$ per madre separata/divorziata);
- lo stato civile del padre ($z_{11} = 1$ per padre sposato/convivente, $z_{12} = 1$ per padre separato/divorziato);
- il titolo massimo conseguito dai genitori ($z_{13} = 1$ per titolo medio, $z_{14} = 1$ per titolo alto);
- l'occupazione della madre ($z_{15} = 1$ per madre part-time, $z_{16} = 1$ per madre casalinga, $z_{17} = 1$ per madre disoccupata/studente);
- l'occupazione del padre ($z_{18} = 1$ per padre part-time, $z_{19} = 1$ per padre disoccupato/studente);
- il numero di figli (compreso il soggetto) ($z_{20} = 1$ per 1 figlio, $z_{21} = 1$ per 2 o più figli);
- il tipo di raccolta dei dati ($z_{22} = 1$ per raccolta tramite questionario);
- il luogo di raccolta ($z_{23} = 1$ per conoscenze personali, $z_{24} = 1$ per lavoro, $z_{25} = 1$ per sede sportiva, $z_{26} = 1$ per scuola);
- l'intervistatore (da z_{27} a z_{54}, tutte variabili dummy per l'intervistatore).

Essendo tutte variabili qualitative a eccezione dell'età dei soggetti, le covariate vengono codificate in variabili *dummy*, inserendole quindi nel modello di regressione lineare come valori binari z_i (assumono 0 o 1) in modo da renderle più facilmente interpretabili rispetto alla categoria di riferimento. Ciascuna covariata qualitativa di k modalità viene pertanto espressa tramite $k - 1$ variabili dummy.
Per l'adattamento e la selezione del modello è stato utilizzato un approccio *stepwise selection*, una combinazione delle due procedure *backward elimination* e *forward selection*. A partire dal modello di regressione lineare che include tutte le variabili a disposizione, ovvero:

$$y = \alpha + \gamma_1 z_1 + ... + \gamma_{54} z_{54} + \epsilon = \alpha + \sum_{i=1}^{54} \gamma_i z_i + \epsilon \quad ,$$

tale approccio utilizza il criterio AIC per selezionare il modello che minimizza la perdita di informazione, ovvero quello con valore minore di AIC. La selezione delle covariate da includere nel modello avviene come nella *forward selection*; dopo l'inserimento di ciascuna variabile, il modello viene riconsiderato per verificare se vi è qualche variabile da eliminare sulla base dell'AIC ottenuto per ciascun modello (come nella *backward elimination*).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
eta_in_mesi	1.000	937.317	937.317	4058.405	0.000
genere	1.000	2.165	2.165	9.373	0.002
naz_genitori	3.000	4.571	1.524	6.597	0.000
titolo_max	2.000	2.522	1.261	5.459	0.004
lavoro_padre	2.000	2.498	1.249	5.409	0.005
interv	28.000	98.065	3.502	15.164	0.000
Residuals	1430.000	330.268	0.231		
R2	**0.760**				
R2 adj.	**0.754**				

Tabella 4.1: Tabella ANOVA del modello di regressione lineare risultante dall'ottimizzazione dell'AIC, con relativo indice R^2.

Con i dati a disposizione nel dataset il modello iniziale contiene tutte le 14 variabili esplicative elencate precedentemente. Le variabili qualitative con 2 o più modalità vengono codificate in variabili dummy ottenendo un modello con 54 regressori.
Utilizzando la procedura *stepwise selection* con il criterio AIC si seleziona il modello che si adatta meglio ai dati e che riduce al minimo la perdita d'informazione, ovvero quello che ha i seguenti regressori: l'età e il genere dei soggetti, la nazionalità e il titolo massimo conseguito dai genitori, l'occupazione del padre e l'intervistatore. Dall'output del modello si nota che tali variabili risultano essere significative, infatti si rifiuta l'ipotesi che i coefficienti stimati per ciascun regressore siano uguali a 0 a un livello di significatività che varia dall'1% al 10%. Come si evince dalla Tabella 4.1, i p-value dei test F calcolati per ciascuna esplicativa indicano che le differenze tra i gruppi sono statisticamente significative.
L'indice R^2 ottenuto dall'output di R è pari circa a 0,76 e l'indice R^2 corretto è 0,754, quindi la varianza spiegata dalle variabili indipendenti del modello selezionato risulta essere pari circa al 76% della varianza della variabile risposta. Alla statistica F per la significatività del modello viene associato un p-value praticamente nullo, pertanto il modello così descritto sembra avere buone capacità predittive.
Successivamente si costruirà un modello con le covariate risultate significative nell'analisi di regressione lineare, integrando però anche la variabile relativa alla concentrazione media mensile di PM_{10} durante la gravidanza, considerata come dato funzionale. Tale modello si indica come modello di regressione lineare con un dato funzionale come regressore.

4.4 Il PM_{10} come dato funzionale

Come abbiamo visto nel paragrafo 4.2, i dati funzionali sono costruiti a partire da sistemi di basi ϕ_k, ovvero sono una loro combinazione lineare:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = c'\phi(t)$$

Il primo passo è costruire funzioni di base opportune per i dati a disposizione. Le concentrazioni di PM_{10} hanno un andamento che non si può descrivere con una funzione periodica, pertanto si utilizzano le *B-splines* essendo le più adeguate in tale contesto. Una volta selezionata la base opportuna vengono stimati i coefficienti c_k relativi alle funzioni di base ϕ_k che, combinate opportunamente tra loro, realizzeranno il dato funzionale.

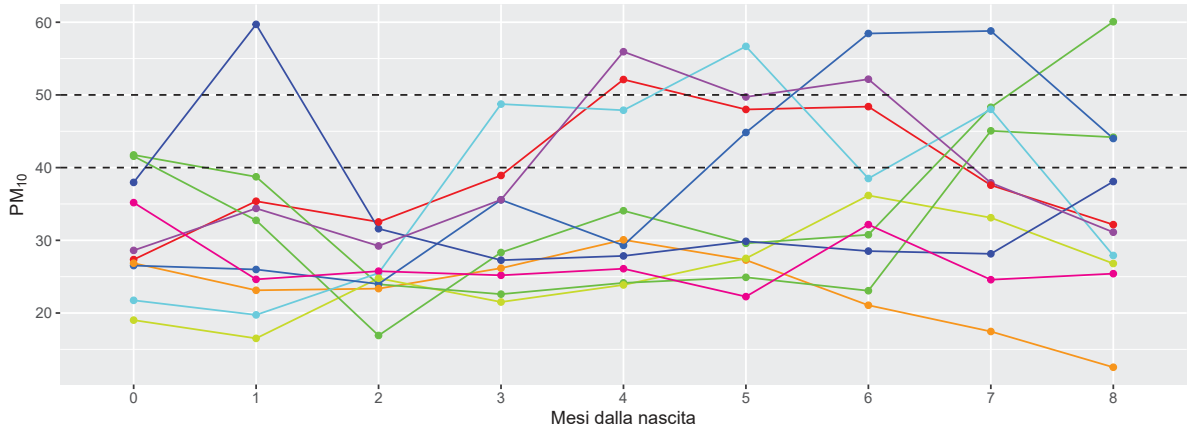

Figura 4.1: Concentrazioni medie di PM_{10} al variare dei mesi precedenti alla nascita per 10 soggetti estratti casualmente dal dataset.

Dalla Figura 4.1 si possono osservare le concentrazioni medie mensili per ciascuno dei 10 soggetti estratti casualmente dal dataset analizzato. Per evidenziare un possibile andamento delle concentrazioni di particolato durante la gravidanza viene delineata una spezzata che unisce ogni osservazione.
Vengono costruite 4 funzioni di base, numero ritenuto sufficiente per rappresentare l'andamento temporale delle 9 osservazioni, tramite la libreria specifica per i dati funzionali (fda). Esse sono descritte in Figura 4.2.
Per ciascun soggetto la variabile funzionale si indica quindi in questo modo, dove $K = 4$:

$$\hat{x}_i(t) = \sum_{k=1}^{K} \hat{c}_{ik} \phi_k(t) + \epsilon_i$$

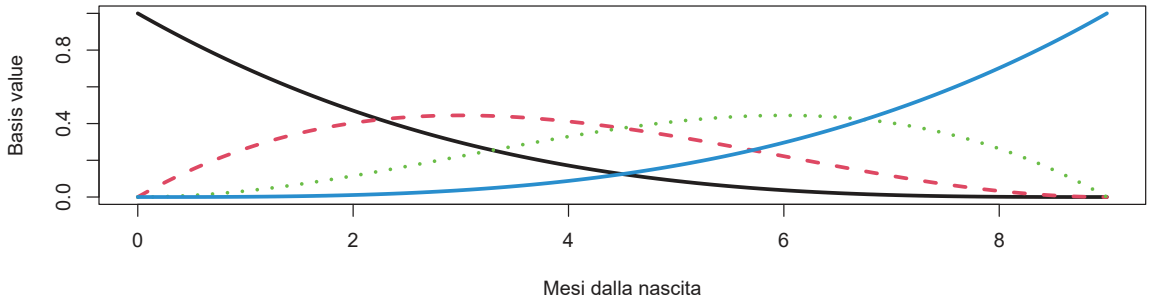

Figura 4.2: Basi di 4 B-splines

Attraverso le basi *B-splines* così costruite, si è potuta stimare la concentrazione di PM_{10} durante la gestazione per ciascun bambino con una curva liscia, in modo tale da visualizzare come cresce o decresce il livello di inquinamento nel tempo.
Dalla Figura 4.3 si osservano le stime delle concentrazioni medie di particolato ottenute al variare dei mesi precedenti alla nascita per i 10 bambini. Si deduce che ciascun soggetto è esposto al PM_{10} in maniera diversa, poiché tali dati dipendono dalla provenienza geografica di ogni bambino. Si può osservare che il soggetto individuato dalla curva rossa, ad esempio, è esposto a una maggior concentrazione di PM_{10} al primo e all'ultimo mese di gravidanza, anche se al concepimento l'esposizione è più ridotta. Per quanto riguarda il soggetto rappresentato dalla curva gialla, invece, si nota un aumento dell'esposizione al PM_{10} nel periodo che va tra il 2° e il 4° mese di gravidanza. Il soggetto identificato dalla curva rosa mostra un'esposizione inferiore al PM_{10} rispetto agli altri bambini considerati, con una maggiore esposizione nei primi mesi della gestazione.
In generale si osserva che alcuni bambini sono esposti all'inquinamento in maniera significativa, infatti si nota che in alcuni mesi della gravidanza la concentrazione media mensile di PM_{10} supera il limite massimo giornaliero consentito di 50 $\mu g/m^3$.

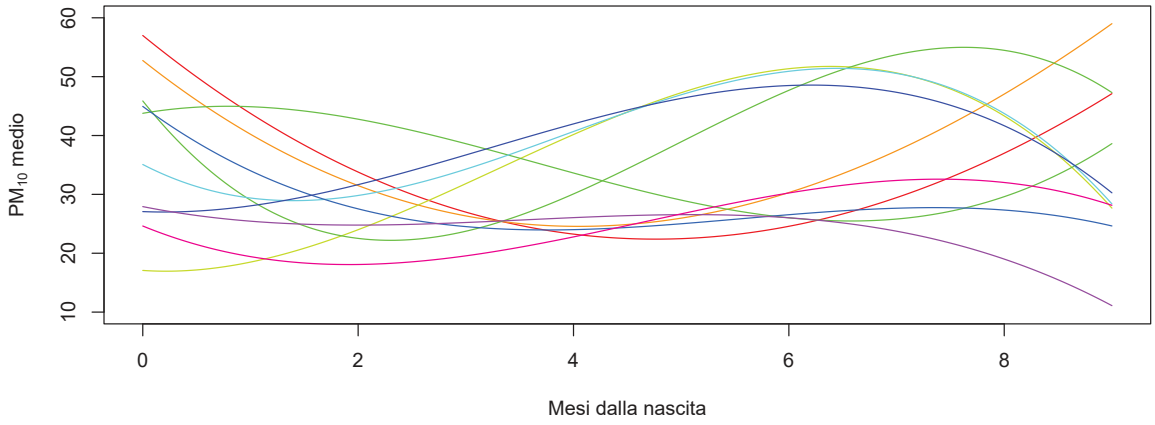

Figura 4.3: Concentrazioni medie di PM_{10} al variare dei mesi precedenti alla nascita stimate attraverso le B-splines per 10 bambini dello studio.

4.5 Il modello

Per questa tesi si è scelto di ricorrere a un modello di regressione lineare avente un termine funzionale tra i regressori. Considerando la variabile risposta y_i di tipo scalare e le covariate di tipo funzionale in un dominio temporale $[0, T]$, basterà sostituire i coefficienti β delle x_i con una funzione $\beta(\cdot)$. Le altre variabili entrano a far parte del modello come confondenti $(\boldsymbol{z}_i)$ con coefficienti $\boldsymbol{\gamma}$. Il modello sarà quindi espresso nel seguente modo:

$$y_i = \alpha + \int_0^T \beta(t)x_i(t)dt + \sum_j \gamma_j z_{ij} + \epsilon_i = \alpha + \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \boldsymbol{\gamma}^T \boldsymbol{z}_i + \epsilon_i$$

dove

$$x_i(t) = \sum_{k=1}^{K} \hat{c}_{ik}\phi_k(t) + \epsilon_i.$$

Nella regressione lineare il calcolo del parametro β si basa sulla minimizzazione della somma degli scarti al quadrato, e si cerca di utilizzare lo stesso anche per quanto riguarda l'approccio funzionale, prestando attenzione alle peculiarità dei dati funzionali. Il criterio per la stima dei parametri è

$$LMSSE(\alpha, \beta, \gamma) = \sum_{i=1}^{N}(y_i - \alpha - \langle \boldsymbol{\beta}, x_i \rangle - \boldsymbol{\gamma}^T \boldsymbol{z}_i)^2 = ||\boldsymbol{y} - \alpha - \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle - \boldsymbol{\gamma}^T \boldsymbol{z}_i||^2$$

dove $\boldsymbol{x}$ rappresenta il vettore delle funzioni covariate $(x_i, ..., x_N)^T$, mentre l'operatore $||\cdot||$ indica la norma euclidea del vettore calcolato.

In particolare, con i dati a disposizione, la variabile risposta è lo score, rappresentante lo sviluppo del soggetto come si è visto nel Capitolo 3, e la variabile inclusa sotto forma funzionale è la covariata che rappresenta le concentrazioni di PM_{10} durante il periodo di gravidanza delle madri dei soggetti.
Tra le variabili esplicative analizzate fino a questo momento, vengono incluse nel modello come variabili confondenti quelle selezionate nel paragrafo 4.3. Tra queste emerge certamente l'età dei soggetti, espressa in forma numerica continua, per la quale si è osservata una correlazione elevata con la variabile di interesse. Dato che dalle analisi precedenti si è riscontrata una relazione di tipo cubico tra lo sviluppo e l'età dei bambini, si verifica se anche il quadrato e il cubo della variabile età (in mesi) possano essere significativamente correlate con la variabile di interesse. Innanzitutto si considera il modello lineare per la selezione della variabili confondenti, aggiungendo il quadrato e il cubo della covariata *eta_in_mesi*. Risulta che esse sono significative e, pertanto, si possono includere nel modello lineare con il termine funzionale tra le esplicative.
Le altre covariate che si includono nel modello sono il genere, alcune caratteristiche relative ai genitori come la nazionalità, il titolo di studio massimo conseguito e l'occupazione del padre, e infine la variabile indicante il rilevatore che ha raccolto i dati.

Dalla Figura 4.4 sottostante si può osservare l'effetto che ha l'inquinamento atmosferico sullo sviluppo dei bambini dello studio durante il periodo di gravidanza con le relative bande di confidenza, al netto di tutte le variabili confondenti. Tale effetto è di tipo funzionale, pertanto $\beta(t)$ è un coefficiente funzionale e la sua stima si esprime:

$$\hat{\beta}(t) = \sum_{k=1}^{K} \hat{\beta}_k \phi_k(t) = \hat{\beta}^T \phi(t)$$

A partire all'incirca dai 6 mesi fino ai 2 mesi prima della nascita, quindi dal 3° all'8° mese di gestazione, questo effetto risulta essere negativo, ovvero sembra influenzare negativamente lo sviluppo del bambino. Osservando più attentamente, le bande di confidenza si trovano al di sotto dello 0 tra il 4° e il 6° mese di gravidanza, quindi concentrazioni elevate di PM_{10} in questi mesi costituiscono un deficit significativo dello sviluppo del bambino.

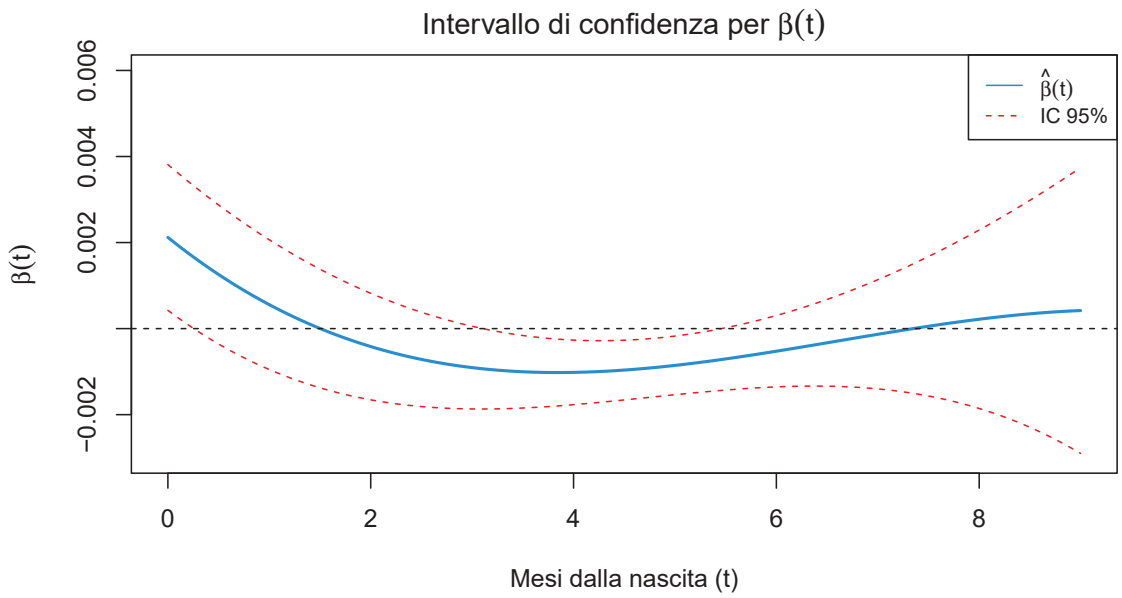

Figura 4.4: Stima dell'effetto dell'inquinamento di PM_{10} sullo sviluppo del bambino durante la gravidanza, con relative bande di confidenza al 95%.

Il modello di regressione lineare con la concentrazione di PM_{10} in forma funzionale, ottenuto dalle analisi svolte, mostra che le variabili esplicative selezionate sono significative, ovvero influiscono sullo sviluppo dei bambini appartenenti allo studio. In Tabella 4.2 vengono riportate le stime dei coefficienti di regressione affiancate dal relativo intervallo di confidenza.
Per quelle stime per cui l'intervallo di confidenza relativo comprende lo 0, si ha che il gruppo di individui identificato da quel regressore non è significativamente diverso dal gruppo di riferimento. Per il modello considerato, infatti, si ha che lo sviluppo dei soggetti aventi entrambi i genitori italiani risulta essere significativamente diverso soltanto da quello dei bambini con i genitori stranieri, ma non diverso dallo sviluppo dei bambini con solo un genitore di nazionalità italiana. Allo stesso modo, lo sviluppo dei soggetti che hanno padre disoccupato/studente oppure part-time non è significativamente differente rispetto a quello dei soggetti con padre full-time.
Considerando la variabile relativa all'età dei soggetti, si osserva che incrementando di 1 mese l'età il punteggio attribuito allo sviluppo del bambino aumenta di 0,35 circa, a parità di tutte le altre variabili. A proposito del genere del bambino, invece, si ha che le femmine hanno uno sviluppo superiore di 0,007 "punti" rispetto ai maschi, sempre al netto delle altre esplicative. Per quanto riguarda la variabile indicante il rilevatore dei dati, si osserva che in Tabella 4.2 ne vengono riportati solamente alcuni livelli, per rendere più leggibile l'output ottenuto.

Variabile	Stima	IC 95%
intercetta	-0.905	[-1.157 , -0.654]
eta	0.349	[0.203 , 0.494]
eta^2	-0.049	[-0.077 , -0.022]
eta^3	0.0029	[0.0012 , 0.0046]
genere Femmina	0.007	[0.002 , 0.013]
genitori Ita Non-Ita	0.019	[0.004 , 0.035]
genitori Non-Ita Ita	-0.007	[-0.027 , 0.013]
genitori Non-Ita Non-Ita	0.017	[-0.015 , 0.049]
titolo medio	-0.006	[-0.012 , -0.001]
titolo alto	-0.019	[-0.036 , -0.003]
padre part-time	0.002	[-0.013 , 0.016]
padre disoccupato/studente	-0.011	[-0.048 , 0.025]
interv.3	-0.041	[-0.076 , -0.006]
interv.8	-0.023	[-0.04 , -0.005]
interv.12	-0.046	[-0.083 , -0.009]

Tabella 4.2: Stime e intervalli di confidenza dei coefficienti di regressione del modello di regressione lineare con il PM_{10} come dato funzionale fra i regressori.

Si vuole studiare se il modello di regressione lineare descritto in questo paragrafo si adatta bene ai dati. La parte di variabilità che non è spiegata dal modello costituisce i residui della regressione e, infatti, essi rappresentano le differenze tra i valori osservati della variabile risposta e i valori stimati con il modello di regressione lineare.

Per verificare se il modello ha una buona capacità predittiva, si esegue un'analisi sui residui. Tramite il test di Shapiro-Wilk si ottiene che i residui non seguono una distribuzione normale, rifiutando l'ipotesi di normalità a tutti i livelli di significatività usuali (p-value $< 0{,}01$). Analogamente, dal Q-Q plot rappresentato in Figura 4.5 si può osservare che i punti non si dispongono lungo la bisettrice, ovvero i quantili dei residui non coincidono perfettamente con quelli della distribuzione normale. Tuttavia, vediamo che solamente in corrispondenza della coda inferiore vi è un discostamento dalla bisettrice e questo probabilmente è dovuto alla presenza di molti outliers in quell'intervallo.

Sempre in Figura 4.5 viene riportato il diagramma di dispersione dei residui rispetto ai valori predetti dal modello. Tale grafico permette di verificare se la distribuzione dei residui sia lineare e se i residui siano omoschedastici. Si osserva che non vi è un andamento sistematico dei residui a eccezione della parte iniziale dove l'andamento non è lineare, pertanto i residui sembrano rispettare l'ipotesi di linearità se non si considerano gli outliers in corrispondenza dei valori più piccoli dello score. D'altro canto, sembra esserci omogeneità della varianza dei residui poiché i punti sono dispersi in modo abbastanza simile sia nella parte sinistra che in quella destra del grafico.

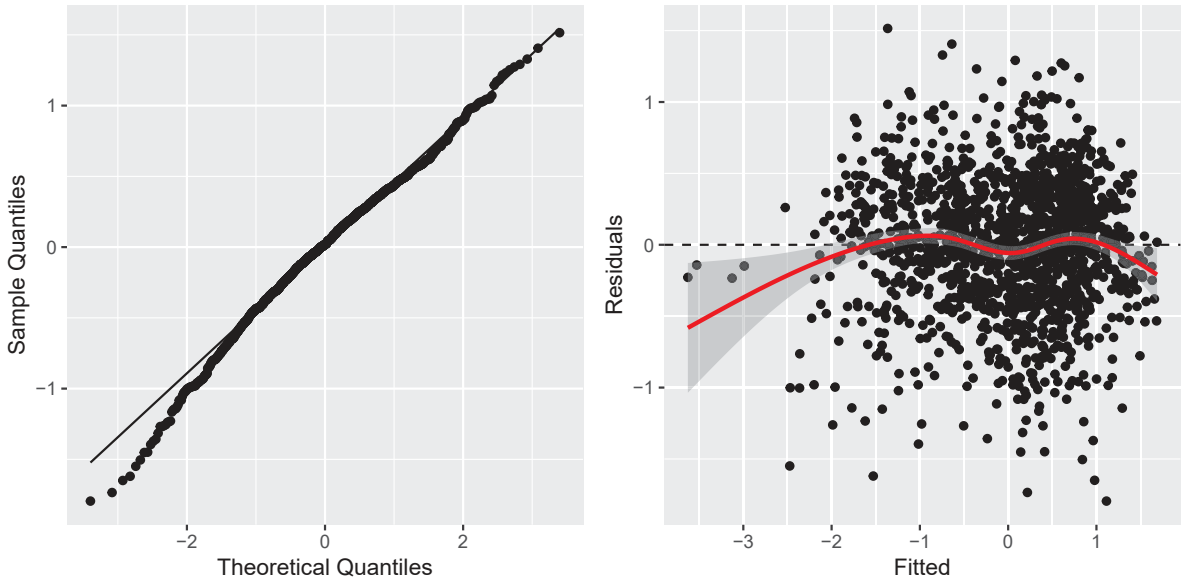

Figura 4.5: Q-Q plot dei residui e diagramma di dispersione dei residui rispetto ai valori predetti del modello di regressione lineare con il PM_{10} sotto forma di dato funzionale.

Per capire quanto è forte la capacità predittiva del modello di regressione costruito si fa ricorso all'indice di bontà dell'adattamento R^2. Mentre per i modelli di regressione lineare classici tale indice viene riportato nell'output di R, per il modello utilizzato in questa tesi bisogna calcolarlo "a mano" a partire dalla sua formula teorica. L'indice R^2, detto anche coefficiente di determinazione, valuta quanta differenza c'è tra i valori osservati della variabile risposta y (score) nel campione e i valori stimati per y dal modello. Piccole discrepanze tra i valori attesi e i valori osservati indicano che il modello si adatta bene ai dati. Al contrario, grandi discrepanze tra questi valori indicano che il modello non spiega bene la variabilità presente nei dati. Il coefficiente di determinazione assume valori compresi tra 0 e 1: più è grande il valore dell'R^2, più il modello ha un alto potere predittivo.
In termini statistici l'R^2 è dato dalla frazione della varianza campionaria di y_i predetta dai regressori x_i. Il coefficiente di determinazione viene definito:

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

dove SQR è la somma spiegata dei quadrati, data dalla somma delle differenze tra i valori predetti di y e la media della stessa variabile dipendente, e SQT è la somma totale dei quadrati, data dalla somma delle differenze tra i valori originari di y e la media della stessa variabile. L'R^2 così calcolato risulta essere pari a 0,764, valore di poco superiore al coefficiente di determinazione ottenuto dal modello di regressione lineare multipla per la selezione delle variabili confondenti. Viene calcolato anche l'R^2 corretto, poiché il coefficiente

di determinazione nella forma "corretta" consente di avere una bontà di adattamento del modello che sia non inflazionata dal numero dei regressori utilizzati ($p = 40$). Tale evidenza consegue dalla stessa formulazione matematica:

$$\tilde{R}^2 = 1 - \frac{SQE/(n-p)}{SQT/(n-1)}$$

dove SQE è la somma dei quadrati dei residui, n è la numerosità campionaria e p è il numero di regressori, che per il modello individuato con il PM_{10} come dato funzionale tra le esplicative è pari a 40. L'indice $\tilde{R}^2$ risulta essere pari a 0,757.

Il modello di regressione lineare avente tra le covariate la concentrazione di PM_{10} sotto forma di dato funzionale e come confondenti le variabili selezionate al paragrafo 4.3 spiega una buona parte (circa il 76%) della varianza totale della variabile risposta, ovvero lo sviluppo dei bambini nello studio. Tuttavia i residui ottenuti da tale modello non verificano l'ipotesi di normalità e non sembrano rispettare la linearità in corrispondenza dei valori inferiori, ma sembrano essere omoschedastici.

Si vogliono confrontare i due modelli analizzati in questo capitolo, ovvero il modello di regressione lineare avente come covariate le variabili confondenti selezionate e il modello lineare con il PM_{10} come dato funzionale tra le covariate e le stesse variabili confondenti del primo. I due modelli differiscono, dunque, solamente del dato funzionale. Si utilizza il test di log-verosimiglianza seguente:

$$LRT = 2 * (l_2(\beta, \gamma; x) - l_1(\gamma; x)) \quad \dot{\sim} \quad \chi^2(4)$$

Il test si esprime come la differenza tra le log-verosimiglianze dei due modelli e va confrontato con la distribuzione asintotica di un Chi-quadro con 4 gradi di libertà, poiché il dato funzionale è costituito da 4 basi B-spline.

Il valore osservato del test LRT e il p-value ad esso associato risultano essere:

$$LRT_{oss} = 19,37 \qquad p = 0,0007$$

dunque, essendo p inferiore al livello di significatività $\alpha = 0,05$, l'inclusione del dato funzionale relativo al PM_{10} è statisticamente significativa.

5 Conclusione

Per questa tesi si è indagato se e quanto lo sviluppo dei bambini possa essere influenzato dall'inquinamento atmosferico causato dal PM_{10} durante la gravidanza. In particolare, gli indici di sviluppo analizzati sono distinti in motorio, adattivo, socioemotivo, cognitivo e comunicativo.
A partire da un campione di 1468 bambini nati fra il 2006 e il 2011, di età compresa tra i 3 gli 8 anni circa, si è potuto osservare che le distribuzioni degli indici di sviluppo associati ai soggetti dello studio presentano diversi outliers in corrispondenza dei valori inferiori. Tale esito è dovuto al fatto che alcuni bambini hanno uno sviluppo piuttosto al di sotto della media.

Nelle analisi bivariate si sono studiate le relazioni tra ciascuna variabile nel dataset e gli indici di sviluppo dei soggetti appartenenti allo studio. Ne è emerso che, marginalmente, le associazioni risultate significative sono quelle con l'età e la provenienza geografica dei bambini, il titolo massimo conseguito dai genitori, l'occupazione del padre, il numero di figli, il luogo di raccolta e il rilevatore dei dati.
Tra queste spicca certamente l'età dei soggetti, infatti è del tutto naturale che lo sviluppo di un bambino, sia questo motorio, cognitivo o adattivo, accresca sempre di più all'aumentare dell'età.

Prima di procedere con la realizzazione di un modello di regressione per i dati a disposizione, si è opportunamente costruita una variabile che sintetizza gli indici di sviluppo, in modo tale da avere un'unica variabile risposta. Tramite un'analisi dei fattori si ottiene quello che viene indicato con il termine *score*, ovvero lo sviluppo dei soggetti, derivante dalla combinazione lineare degli *score regression*. Marginalmente non si osserva una relazione significativa tra la concentrazione di PM_{10} e lo score.

Le concentrazioni di PM_{10} sono state registrate a ogni mese di gravidanza, pertanto si hanno a disposizione 9 osservazioni nel tempo di tale variabile. Data la sua natura, si è scelto di rappresentare la concentrazione di PM_{10} sotto forma di dato funzionale, costruito a partire da sistemi di basi B-splines, le più adeguate in tale contesto.
Successivamente vengono selezionate le variabili confondenti attraverso un modello di regressione lineare avente come variabile dipendente lo sviluppo dei soggetti dello studio (score) e come covariate tutte le altre variabili del dataset, a eccezione delle concentrazioni di PM_{10}. Le variabili confondenti risultano essere: l'età e il genere dei soggetti, la nazionalità dei genitori, il titolo massimo conseguito dai genitori, l'occupazione del padre e l'intervistatore.

Infine si considera un modello di regressione lineare avente la concentrazione di PM_{10} come dato funzionale tra i regressori e le variabili selezionate precedentemente come confondenti. Si osserva che l'effetto dell'inquinamento atmosferico sullo sviluppo dei bambini è significativamente negativo tra il 4° e il 6° mese di gravidanza. Questo significa che a maggiori

concentrazioni di PM_{10} rilevate durante tale periodo corrisponde uno sviluppo inferiore del bambino.

I risultati ottenuti sono in linea con quanto descritto in letteratura, infatti è stato provato che maggiori concentrazioni di inquinamento durante il 4° e il 5° mese della gestazione, mesi in cui si sviluppano gli organi principali del feto tra cui il cervello, hanno un'influenza negativa significativa sullo sviluppo dei bambini.

Appendice

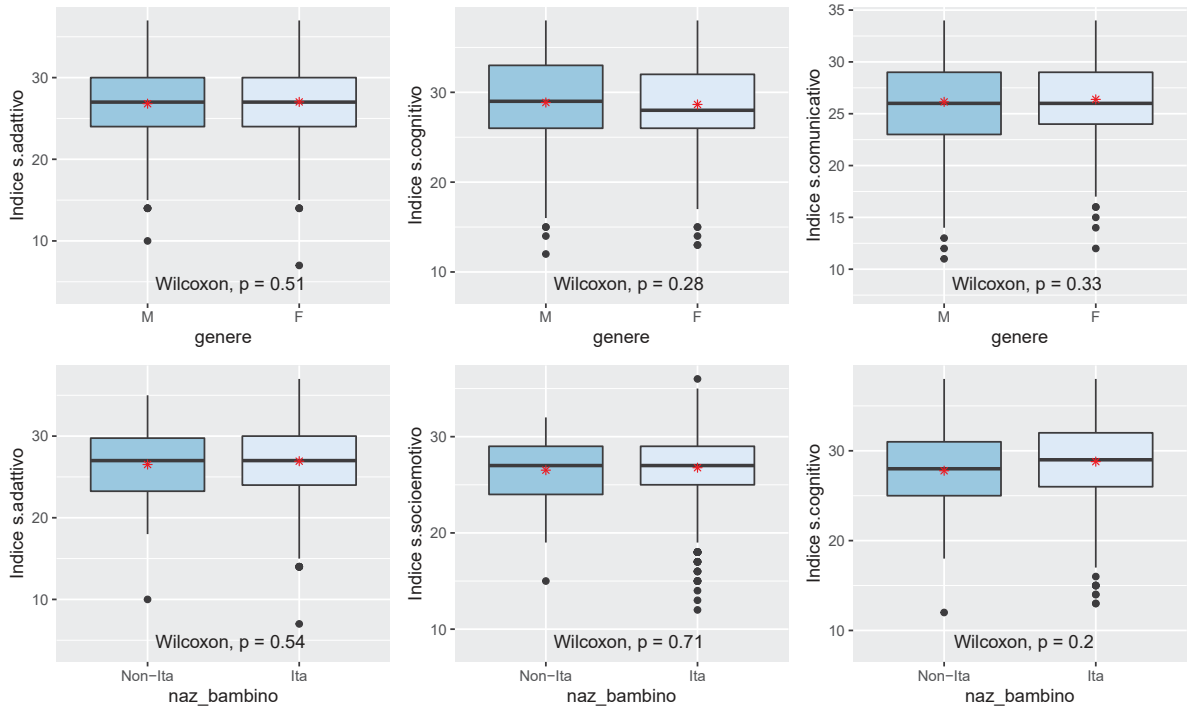

Figura A2.13: Distribuzioni degli indici di sviluppo adattivo, cognitivo e comunicativo dei soggetti a seconda del genere e degli indici di sviluppo adattivo, socioemotivo e cognitivo a seconda della nazionalità dei soggetti.

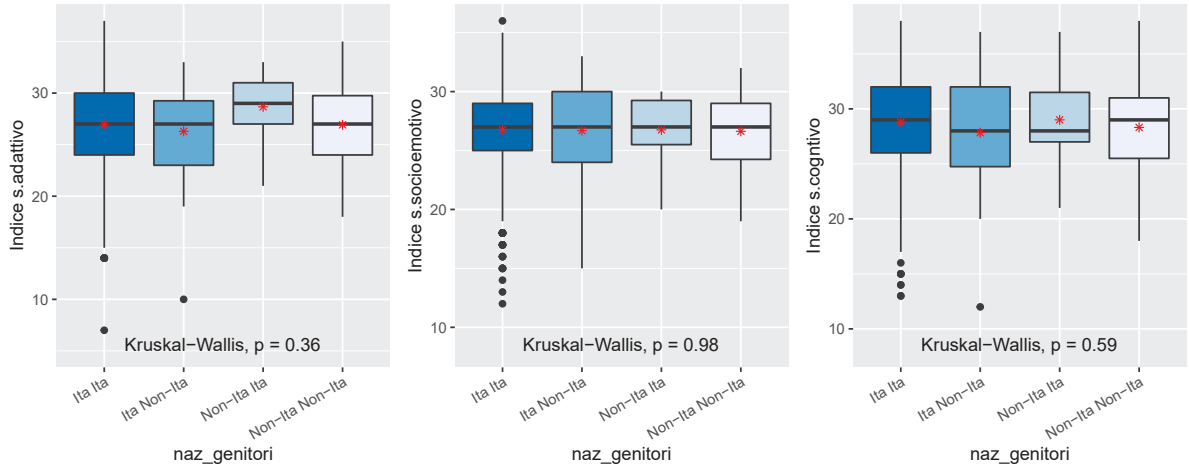

Figura A2.14: Distribuzioni degli indici di sviluppo adattivo, socioemotivo e cognitivo dei soggetti a seconda della nazionalità dei genitori.

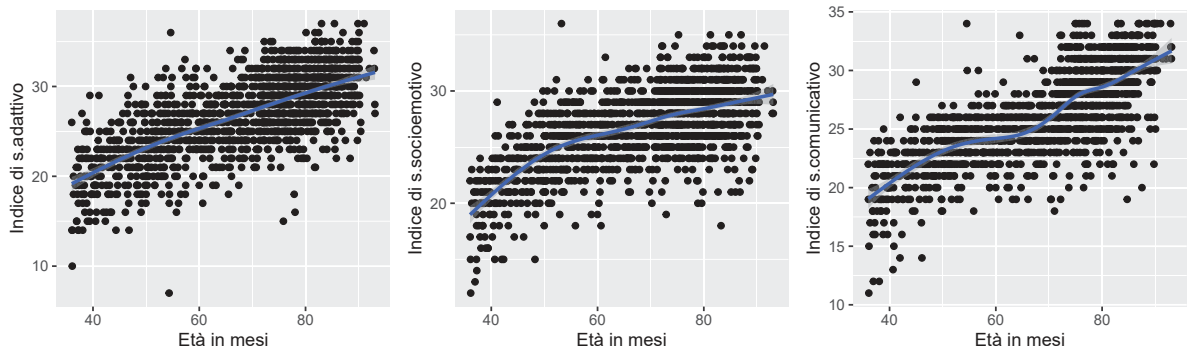

Figura A2.16: Distribuzioni degli indici di sviluppo adattivo, socioemotivo e comunicativo a seconda dell'età dei soggetti nello studio.

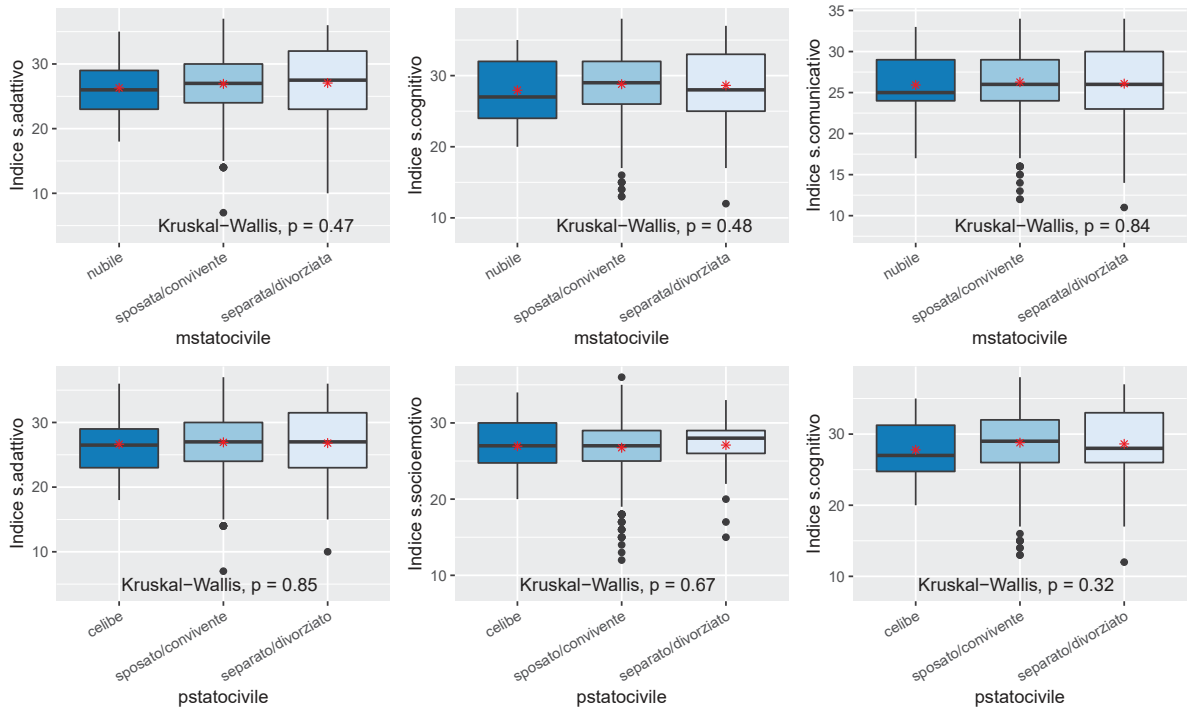

Figura A2.17: Distribuzioni degli indici di sviluppo adattivo, cognitivo e comunicativo a seconda dello stato civile della madre e degli indici di sviluppo adattivo, socioemtivo e cognitivo a seconda dello stato civile del padre dei soggetti nello studio.

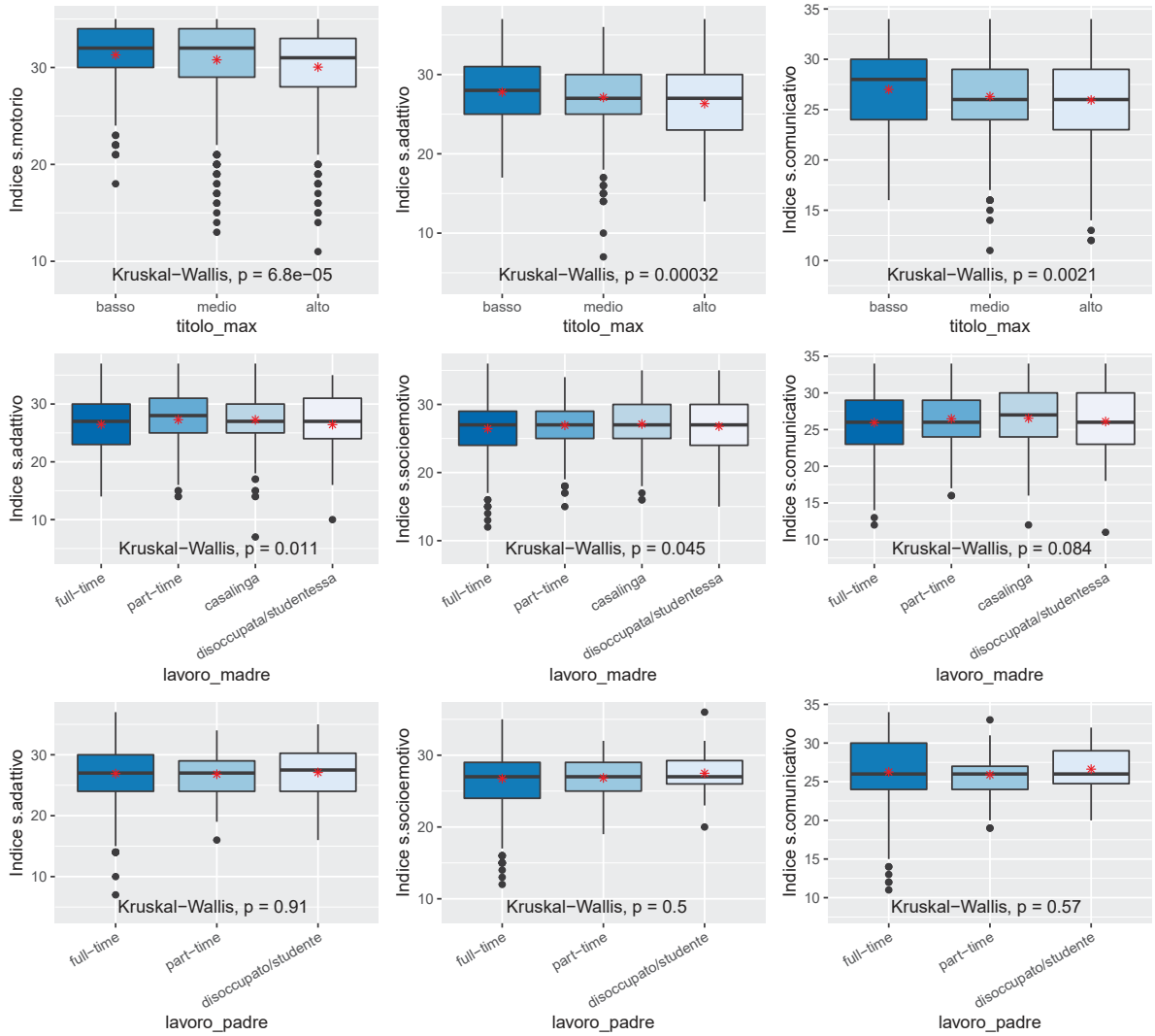

Figura A2.18: Distribuzioni degli indici di sviluppo motorio, adattivo e comunicativo a seconda del titolo di studio massimo conseguito dai genitori dei soggetti, e degli indici di sviluppo adattivo, socioemotivo e comunicativo a seconda dell'occupazione lavorativa dei genitori dei soggetti nello studio.

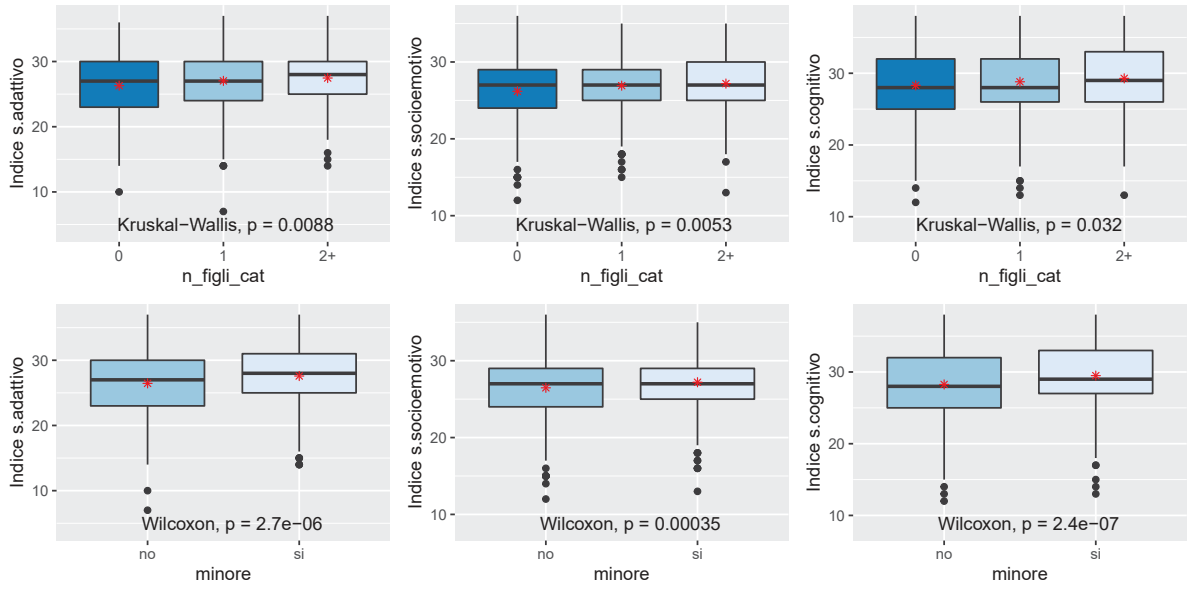

Figura A2.19: Distribuzioni degli indici di sviluppo adattivo, socioemotivo e cognitivo a seconda del numero di figli, compresi i soggetti nello studio, e a seconda che i soggetti abbiano almeno un fratello minore.

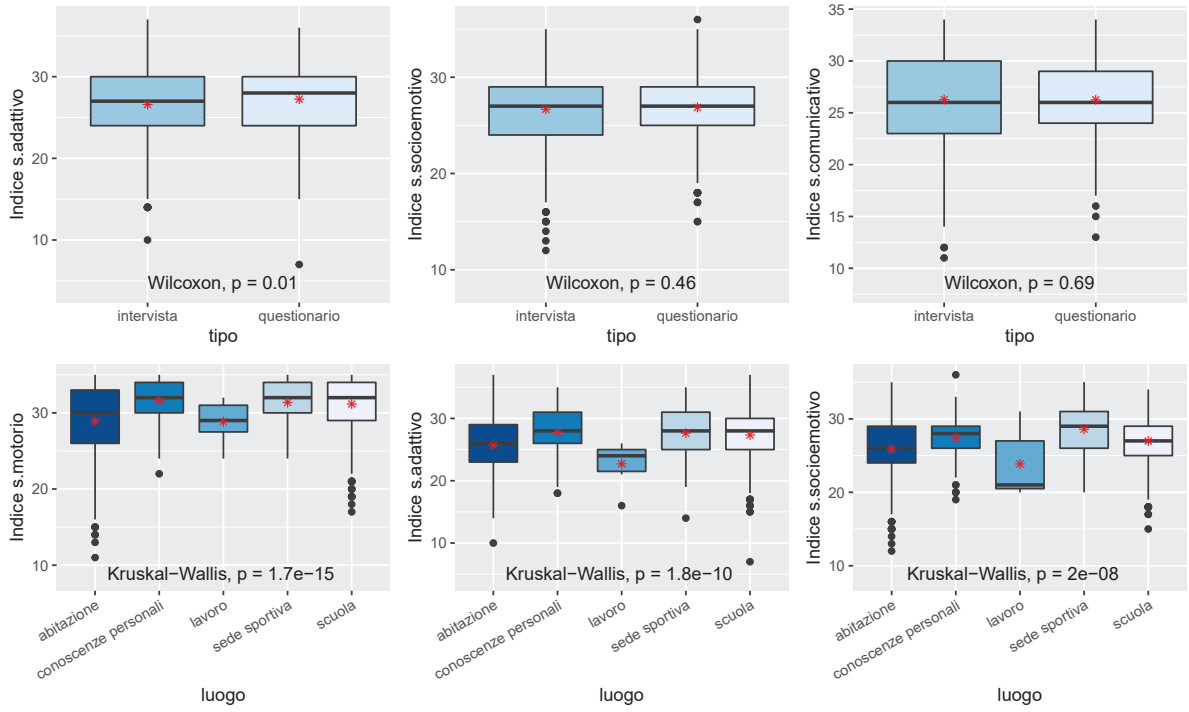

Figura A2.20: Distribuzioni degli indici di sviluppo adattivo, socioemotivo e comunicativo a seconda del tipo di raccolta dei dati, e degli indici motorio, adattivo e socioemotivo a seconda del luogo di raccolta dei dati.

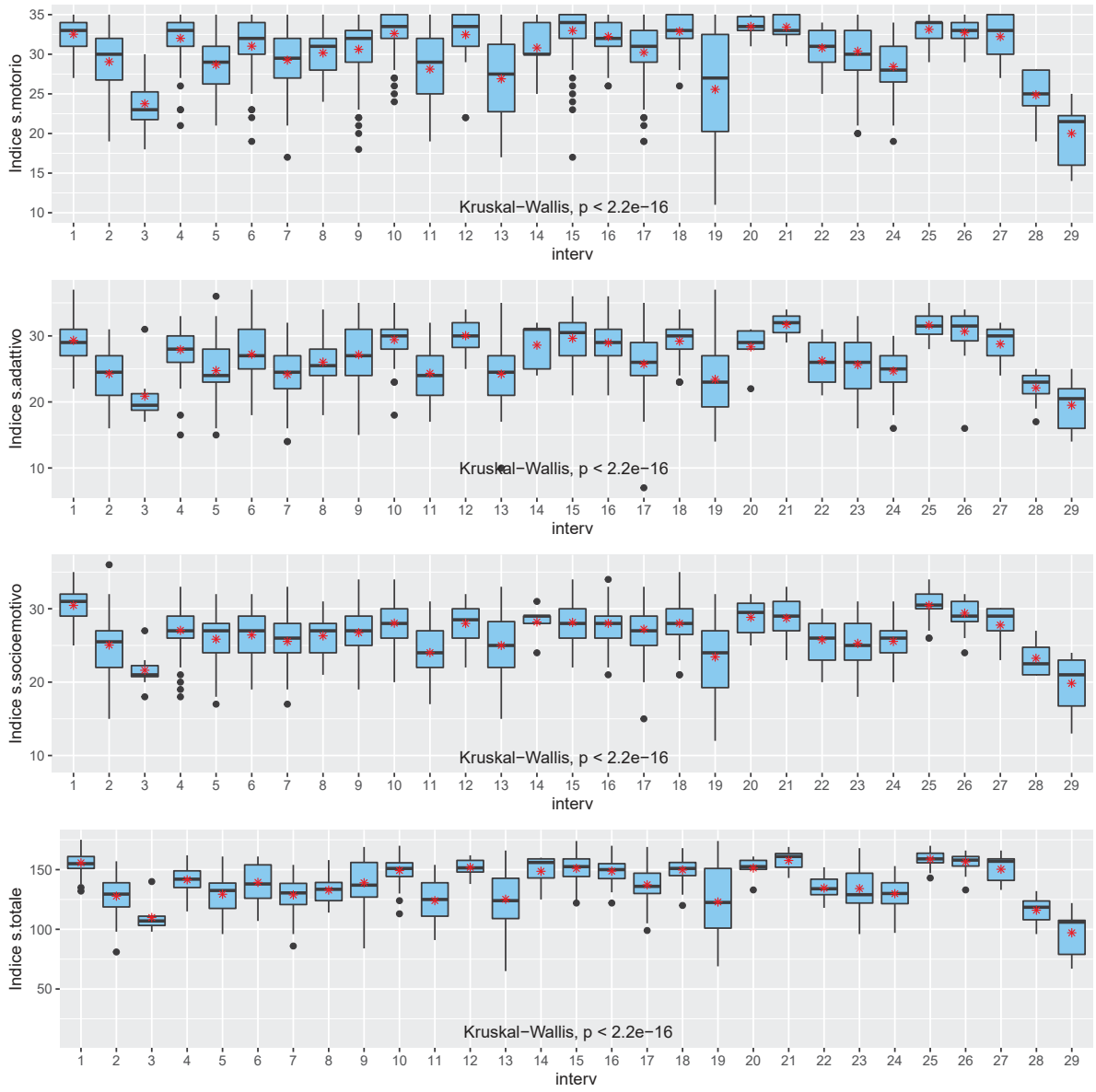

Figura A2.21: Distribuzioni degli indici di sviluppo motorio, adattivo, socioemotivo e totale a seconda della persona che ha rilevato i dati.

Bibliografia

[1] https://www.isprambiente.gov.it/it/attivita/aria-1

[2] https://www.eea.europa.eu/it/themes/air/intro

[3] https://www.focus.it/ambiente/ecologia/inquinamento-potrebbe-avere-effetti-sul-cervello

[4] *Air pollution and cognitive performance in children, APGAR*, Fundacion Privada Instituto De Salud Global Barcelona.
https://cordis.europa.eu/article/id/220360-air-pollution-and-cognitive-performance-in-children

[5] *Particolato*, Ministero della Salute (2015).
https://www.salute.gov.it/imgs/C_17_opuscoliPoster_283_ulteriorallegati_ulteriореallegato_7_alleg.pdf

[6] https://www.inquinamento.org/pm10/pm10-limiti.html

[7] *Analisi dei trend dei principali inquinanti atmosferici in Italia 2003-2012*, ISPRA.
https://www.isprambiente.gov.it/files/pubblicazioni/rapporti/R_203_2014.pdf

[8] https://www.treccani.it/enciclopedia/sviluppo-del-sistema-nervoso-nel-feto_%28Dizionario-di-Medicina%29/

[9] https://toolbox.eupati.eu/resources/la-statistica-negli-studi-clinici-distorsione-bias/?lang=it

[10] http://biometria.univr.it/sesm/files/bias.pdf

[11] *Dispense di Analisi dei Dati Multidimensionali, 2020-2021*, Manuela Cattelan.

[12] Christine DiStefano, Min Zhu, Diana Mindrila, *Understanding and Using Factor Scores: Considerations for the App, 2009*, Practical Assessment, Research, and Evaluation.
https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1226&context=pare