Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

FATTORI CHE INFLUENZANO LA FORMAZIONE DI CORPUSCOLI IN SOGGETTI ESPOSTI AD AMIANTO

Relatore: Prof. Paolo Girardi

Laureando: Piero Trevisan
Matricola N 1162982

Anno Accademico 2019/2020

A tutti coloro che mi hanno sempre sostenuto

Indice

1 Introduzione **3**
1.1 Presentazione del problema . . . 3
1.2 Stato dell'arte . . . 3
1.3 Amianto . . . 4
1.4 Storia dell'amianto in Italia . . . 5
1.5 Malattie legate all'amianto . . . 6
1.6 Obiettivo . . . 7

2 Analisi descrittive **9**
2.1 Descrizione delle caratteristiche . . . 9
2.2 Pulizia del dataset . . . 10
2.2.1 Dati mancanti nel dataset . . . 11
2.2.2 Recupero tramite altre caratteristiche . . . 12
2.2.3 Multivariate Imputation by Chained Equations (MICE) . . . 13
2.2.4 Regressione . . . 14
2.3 Test statistici . . . 19
2.4 Analisi univariate . . . 22
2.5 Analisi bivariate . . . 32

3 Modelli di regressione **43**
3.1 Teoria dei minimi quadrati ordinari . . . 43
3.2 Applicazione del modello OLS . . . 44
3.3 Teoria dei minimi quadrati pesati . . . 48
3.4 Applicazione del modello WLS . . . 50

4 Studio degli outliers **57**
4.1 Introduzione ai valori anomali . . . 57
4.2 Metodo di Farcomeni-Viviani . . . 58
4.3 Applicazione nel modello e risultati . . . 59

5 Conclusioni **65**

Appendice A. Output dei test **66**

Appendice B. Codice dell'algoritmo di Farcomeni-Viviani **70**

Bibliografia **72**

1 Introduzione

Ho svolto questa tesi durante un periodo di stage presso il Servizio Epidemiologico Regionale dell'Azienda Zero; in particolare ho avuto modo di lavorare presso il Registro Mesoteliomi della Regione Veneto. L'esperienza è durata da febbraio 2020 ad agosto 2020 e sono stato seguito dal dott. Ugo Fedeli, responsabile del Registro Mesoteliomi, come tutor aziendale e dal Prof. Paolo Girardi come tutor accademico e come relatore. È stata un'ottima occasione per poter applicare ad un reale problema statistico e medico le conoscenze apprese durante il mio percorso di studi.

1.1 Presentazione del problema

Il tema di questa tesi è l'analisi di un dataset contenente delle misurazioni del carico di amianto polmonare in soggetti esposti ad amianto in passato. Dalla letteratura scientifica si sa che le fibre di amianto vengono inalate e si depositano nei polmoni; queste col passare degli anni si possono trasformare in corpuscoli ed entrambi contribuiscono all'infiammazione locale ed eventualmente a neoplasie maligne. Lo scopo principale del lavoro è stato quello di andare a valutare quali fattori influenzano maggiormente la formazione di corpuscoli, considerati come misura indiretta del danno polmonare e che quindi possono essere correlati alla comparsa di varie malattie. Le misurazioni del numero di fibre e di corpuscoli nel tessuto polmonare sono fondamentali per ricostruire la storia dell'esposizione ad amianto di una persona; in particolar modo il numero di fibre per centimetro cubo per anno esprime un parametro importante per stimare l'esposizione ad amianto. Questo risultato è molto importante poichè è noto [1] che la probabilità di insorgenza di malattie asbesto correlate aumenta quando l'esposizione è prolungata. È inoltre risaputo che al crescere del numero di fibre presenti nel polmone aumenta la formazione di corpuscoli nel corso degli anni e quindi la probabilità di insorgenza di patologie. I dati provengono da analisi di laboratorio dove sono stati analizzati dei campioni di polmone provenienti da biopsie o autopsie in persone con pregressa esposizione ad amianto.

1.2 Stato dell'arte

Il tema del nesso causale tra esposizione ad amianto e malattie è stato ampiamente dibattuto per decenni; si era visto infatti che i lavoratori esposti ad amianto erano più soggetti a patologie come tumori al polmone e mesoteliomi. È invece relativamente più recente lo studio sui fattori che influenzano la formazione dei corpuscoli di amianto. Sono interessanti i risultati ottenuti da uno studio [2] sul rischio di insorgenza del mesotelioma pleurico e del tumore al polmone in relazione all'esposizione ad amianto: esso dipende infatti dalla tipologia di fibra inalata (crisotilo, amosite e crocidolite) e dall'esposizione cumulativa, ovvero la quantità totale di una sostanza (l'amianto) a cui una persona è esposta nel tempo. In uno studio del 1981 [3] vengono analizzate la formazione dei corpuscoli di amianto (e di altri corpi ferruginosi) e la loro correlazione con malattie asbesto correlate. Si nota che

il numero di corpuscoli è un indicatore del numero di fibre lunghe di anfiboli presente nei polmoni analizzati e che è un parziale indicatore di insorgenza del cancro al polmone. Non è stata rilevata invece una correlazione significativa tra il numero di corpuscoli e l'insorgenza di patologie in soggetti che sono stati a contatto con amianto per poco tempo. A parità di tipo di fibre, il numero di corpuscoli che si formano in seguito ad un'esposizione di fibre di crisotilo è più contenuto rispetto ad un'esposizione ad anfiboli. Viene inoltre fatto notare che le donne possono entrare a contatto con amianto anche tramite cosmetici o altri prodotti.
In un lavoro della dott.ssa Somigliana, del dott. Girardi, del dott. Barbieri e del dott. Merler (di natura simile a questa tesi) [4] uno dei principali risultati osservati è la relazione positiva tra numero di fibre di amianto e numero di corpuscoli per tipo di patologia e per tipo di fibra. Inoltre si è scoperto che tale relazione dipende dalle dimensioni della fibra e dal tempo di esposizione: per ogni anno trascorso dalla fine dell'esposizione si stima un aumento del 2.4% del numero di corpuscoli formatisi.
Uno studio del 2002 [5] ha voluto verificare quali sono i fattori correlati con la presenza di corpuscoli su 270 ex lavoratori che sono stati a contatto con amianto, analizzando l'espettorato, un muco che si forma nei polmoni. Si è concluso che i fattori maggiormente correlati sono l'intensità dell'esposizione, il tipo di fibra inalato e gli anni dalla fine dell'esposizione. In particolare aumenta il numero di corpuscoli col passare del tempo e se ne è osservato un numero maggiore in pazienti anziani.
Infine in uno studio del 2004 [6] si è voluta confrontare la presenza di corpuscoli di amianto tra soggetti esposti e non esposti ad amianto tramite l'analisi di un lavaggio broncoalveolare (BAL), un metodo usato per diagnosticare delle patologie dell'apparato respiratorio. Si è notato che un numero maggiore di corpuscoli è collegato a soggetti con esposizione ad amianto, con difficoltà respiratorie come tosse ed altri sintomi.

1.3 Amianto

L'amianto (o asbesto) è un materiale fibroso, costituito da fibre minerali della lunghezza di circa 5µm e con un rapporto lunghezza/diametro di circa 3:1; le piccole dimensioni delle fibre [7] le rendono volatili e possono essere inalate con molta facilità. In particolare con il termine amianto si indicano sei diversi minerali appartenenti alla classe dei silicati. In base alla loro composizione chimica essi sono suddivisi in due gruppi: quelli che contengono calcio e magnesio sono detti anfiboli e comprendono actinolite, amosite (o "amianto bruno"), antofillite, crocidolite (o "amianto blu") e tremolite. Il crisotilo (o "amianto bianco") è invece un silicato di magnesio che appartiene al gruppo del serpentino. Quest'ultimo, tra i vari tipi di amianto, è il più diffuso: si stima infatti che rappresenti oltre il 93% dell'asbesto usato in commercio. Per via della loro diversa struttura chimica, i minerali del gruppo del serpentino vengono chiamati "silicati a fogli ricurvi" poichè i tetraedri di silicio ed ossigeno che li compongono si dispongono formando delle lamelle, mentre quelli del gruppo degli anfiboli sono chiamati "silicati a doppia catena" in quanto i tetraedri formano due catene parallele tra loro.

Una volta cessata l'esposizione, il carico di fibre nel corpo umano può diminuire tramite processi di smaltimento naturale; in particolare avviene la rimozione delle fibre di amianto tramite i macrofagi alveolari. I macrofagi sono un tipo di globuli bianchi del sistema immunitario che hanno la funzione di eliminare le particelle estranee dal nostro organismo, inglobandole tramite un processo chiamato fagocitosi. I macrofagi alveolari [8] risiedono negli alveoli polmonari e rappresentano la prima linea difensiva contro virus, batteri e particelle ambientali inalate. Un altro processo di eliminazione delle fibre è rappresentato dall'espettorato [9], un muco prodotto dai polmoni ed espulso dalla bocca; la sua analisi può essere un metodo non invasivo molto utile per verificare la presenza di amianto nell'organismo. A causa delle loro diverse proprietà chimiche, i tempi di espulsione variano a seconda dei due gruppi, si parla infatti di mesi per le fibre di crisotilo e di anni per quelle di anfiboli.
L'amianto gode di ottime proprietà fisiche che lo rendono tanto usufruibile in numerosi ambiti industriali quanto dannoso per la salute umana. Si tratta infatti di un materiale poco costoso, ad alta flessibilità e facilità di lavorazione, dovute alla sua struttura fibrosa; è inoltre facilmente mescolabile ad altre sostanze come il cemento, dotato di capacità fonoassorbenti e resiste all'usura ed agli agenti biologici e chimici come gli acidi. Sono importantissime anche le proprietà termiche come l'elevata resistenza alla fusione e l'alto grado di isolamento termico; non è infatti un caso che la principale azienda di produzione venne denominata Eternit. -Naturalmente tali caratteristiche possono variare a seconda dei diversi tipi di amianto-.

1.4 Storia dell'amianto in Italia

Come già riportato, l'amianto è stato largamente usato in vari settori industriali [10] (nell'edilizia, nell'industria metalmeccanica e della metallurgia, nei cantieri navali, nell'industria del cemento-amianto e tessile) per le sue ottime proprietà. In Italia nella prima metà del '900 l'amianto cominciò a diffondersi con forza non solo nell'edilizia, ma anche in usi apparentemente inadatti come la costruzione di manufatti di uso quotidiano, ad esempio giocattoli o oggetti di modellistica. Il settore nel quale c'è stato il maggior uso di amianto nel nostro Paese è stato quello dell'edilizia, che nel periodo compreso tra il 1965 e il 1983 ha fatto largo uso del cemento-amianto, chiamato eternit; il materiale era così diffuso ed utilizzato che l'Italia è stata fino alla fine degli anni '80 il primo produttore europeo di amianto. Alcuni classici esempi di utilizzo dell'amianto nell'edilizia furono i cassoni, serbatoi e tubazioni per l'acqua, camini, canne fumarie, pavimentazioni, caldaie, stufe e forni. Nell'industria l'amianto è stato ampiamente utilizzato nella costruzione di pannelli e coperture, nella fabbricazione di tubi coibentati, serbatoi, reattori e refrigeratori, nella costruzione di alcune componenti di macchine e nelle guarnizioni. In seguito alla scoperta di significativi danni alla salute degli esposti, nel 1992 vennero vietate la produzione e l'installazione di materiali contenenti amianto, mentre le bonifiche sono tuttora in corso in tutto il Paese.
L'amianto viene estratto come materia prima da miniere e cave che, nella maggior parte dei

casi, si presentano a cielo aperto; tuttavia esistono alcune miniere in cui il minerale viene estratto in profondità. Solo il crisotilo, l'amosite e la crocidolite hanno avuto nel tempo una notevole importanza nel settore industriale, mentre i rimanenti minerali di amianto sono stati usati saltuariamente. In Italia il 75% della produzione di amianto era destinata alla fabbricazione del fibrocemento mentre il restante 25% era destinato in maniera quasi esclusiva alla fabbricazione di materiali di frizione.

1.5 Malattie legate all'amianto

La pericolosità dell'amianto è legata alla liberazione nell'ambiente delle fibre che lo compongono; quando vengono inalate si depositano nei polmoni ed il loro smaltimento naturale è piuttosto lento. In particolare subiscono processi di trasposizione e bio-trasformazione come la loro trasformazione in corpuscoli; una parte di loro viene infatti inglobata nelle estremità dai macrofagi formando dei corpuscoli di amianto (asbestos bodies). Il processo che avviene prende il nome di fagocitosi frustrata e si verifica quando tantissimi macrofagi tentano inutilmente di inglobare delle fibre più grandi di loro, formando un corpuscolo; questo processo provoca uno stato di infiammazione persistente con conseguente danneggiamento delle cellule epiteliali. È opportuno sottolineare che le fibre con impatto maggiore sulla salute sono quelle lunghe e sottili (con diametro $< 1.5\mu m$ e lunghezza maggiore di $8\mu m$) in quanto sono più difficili da inglobare e per via della loro forma possono penetrare più profondamente nei polmoni.

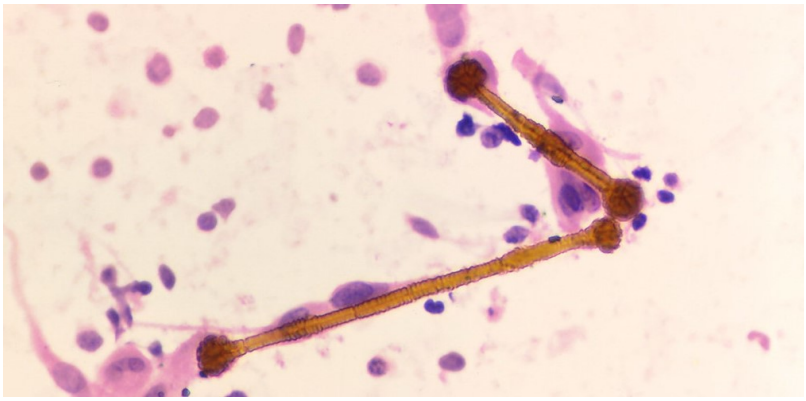

Figura 1.5: Esempio di fibre di amianto inglobate, che formano dei corpuscoli.

Un'importante patologia asbesto correlata è il mesotelioma pleurico [11], un tumore che interessa il mesotelio, un tessuto che riveste come una sottile pellicola la parete interna del torace, dell'addome e dello spazio intorno al cuore. Assume nomi differenti a seconda della zona del corpo colpita e della tipologia di cellula maligna presente nel tumore. Si tratta di un tumore molto raro, con un incidenza per 100000 abitanti in Italia di 3,6 casi negli uomini e 1,6 nelle donne, che cresce se si è esposti ad amianto per molto tempo. Il mesotelioma è piuttosto difficile da diagnosticare poichè presenta sintomi simili ad altre malattie (difficoltà respiratoria, tosse e dolore toracico); in genere ha un tempo di latenza molto lungo, quantificabile in decenni. Non essendo previsti screening per la diagnosi

precoce, la diagnosi può avvenire tramite esami periodici come radiografia o TC.
L'asbestosi [12] è una malattia cronica polmonare provocata da un'esposizione prolungata e intensa alle fibre di amianto. In genere si manifesta dopo 15 anni dall'esposizione ed i sintomi più frequenti sono tosse, insufficienza respiratoria ed alterazione della funzionalità polmonare; è una malattia irreversibile e la terapia può riguardare solo le sue complicanze.
Nonostante la complessa eziologia del tumore al polmone, diversi studi epidemiologici hanno evidenziato che un aumento significativo del suo rischio è associato all'esposizione ad asbesto. Si può sviluppare nelle cellule che formano bronchi, bronchioli ed alveoli e può costituire una massa che blocca il corretto flusso dell'aria, oppure provocare emorragie polmonari o bronchiali. Ne esistono varie tipologie a seconda del tessuto polmonare interessato. I sintomi e la diagnosi sono simili al mesotelioma pleurico.

1.6 Obiettivo

L'obiettivo di questa tesi è analizzare il dataset per valutare quali sono i fattori che più influenzano la formazione di corpuscoli in pazienti esposti ad amianto. Sarà inoltre interessante confrontare i risultati ottenuti con quelli già conosciuti dalla letteratura scientifica. Verranno utilizzati opportuni metodi statistici per valutare se le variabili esplicative a disposizione influiscono o meno sul fenomeno d'interesse.
Nella fase iniziale di analisi in cui viene esaminata la distribuzione della risposta a seconda delle varie esplicative; verranno usate tabelle, grafici esplicativi e test a seconda della natura dei dati. L'analisi congiunta dei fattori sarà invece effettuata da modelli di regressione.
Nell'ultimo capitolo verrà utilizzato un algoritmo specifico per il trattamento dei valori anomali.

2 Analisi descrittive

In questo capitolo viene presentato il database oggetto di lavoro, compreso il suo percorso di pulizia. Contiene i dati riferiti ad un insieme di persone che sono state a contatto con amianto; assieme alla quantità d'interesse ovvero il numero di corpuscoli [13] trovato nei campioni di polmone, è presente una serie di caratteristiche relative alla vita lavorativa dei pazienti ed alla loro esposizione. Queste ultime sono fondamentali per conoscere al meglio le dinamiche della formazione dei corpuscoli.
Il database è stato formato durante un'attività di ricerca promossa dal Registro Regionale Veneto dei casi di Mesotelioma in collaborazione con il Centro di Microscopia Elettronica dell'ARPA di Milano. I pazienti oggetto di studio sono affetti principalmente da patologie asbesto correlate e sono stati sottoposti ad autopsie giudiziarie o ad interventi chirurgici. Dopo la raccolta di ogni campione di tessuto polmonare, il reperto viene spedito alla dott.ssa Somigliana per le analisi di laboratorio, dove il campione conservato sotto formalina subisce un processo di nome "digestione chimica". Per eliminare la parte organica del campione ed avere solo quella inorganica (le fibre di amianto) viene prima applicato un acido e successivamente viene incenerito tramite un bruciatore da laboratorio. Avviene poi la suddivisione del campione in altri sottocampioni chiamati "campi" e tramite un potente microscopio ottico avviene la misurazione del numero di fibre trovate per grammo di tessuto secco. Individuare i corpuscoli è un lavoro più semplice in quanto sono di grandezza superiore, ma per facilitarne l'identificazione è stato predisposto un atlante di corpuscoli di amianto dove per ogni tipologia viene riportata la forma, il colore ed altre caratteristiche come la natura del rivestimento delle fibre.

2.1 Descrizione delle caratteristiche

Il dataset oggetto di lavoro proviene dal Registro Mesoteliomi della Regione Veneto e non è stato precedentemente studiato. Il compito è stato quello di pulirlo ed analizzarlo tramite tecniche statistiche per osservare quali sono le variabili che influiscono maggiormente sulla formazione dei corpuscoli in pazienti esposti ad amianto. Il dataset analizzato contiene 193 soggetti e 26 caratteristiche. Ne viene riportata una breve descrizione, organizzata in gruppi per motivi di praticità.

- *id, genere, fonte, regione*: sono delle caratteristiche demografiche e di residenza come il genere ("D" per donna, "U" per uomo), il luogo da cui è arrivata l'informazione (Gorizia, Padova o altro) e la regione di residenza (Friuli Venezia Giulia, Veneto o altro). *id* è un identificatore che individua il soggetto;
- *malattia, intervento*: sono delle caratteristiche che identificano il paziente dal punto di vista medico. Indicano la patologia di cui soffre il soggetto (mesotelioma, tumore al polmone o altro) ed il tipo di intervento a cui è stato sottoposto (biopsia o autopsia);
- *stato.vita*, *data.nascita*, *data.inizio*, *data.fine*, *data.prelievo*, *data.diagnosi*,

data.analisi.corpuscoli, *data.decesso*: sono delle caratteristiche temporali che descrivono la storia medica di ogni paziente, ovvero le date di nascita, di inizio e fine esposizione, del prelievo del tessuto polmonare, della diagnosi, dell'analisi dei corpuscoli e di morte (se è avvenuta). Per quanto riguarda *stato.vita*, è una caratteristica dicotomica che indica lo stato di follow-up del soggetto: vale 3 se vivo, 6 se deceduto;

- *settore.lavorativo, causa.esposizione, fibra.inalata, livello.esposizione*: è un gruppo di caratteristiche che fa riferimento all'esperienza lavorativa dei soggetti. In particolare *settore.lavorativo* indica il settore lavorativo nel quale l'unità statistica è stata a contatto con amianto (cantieri navali, impianti industriali, mezzi ferroviari o altro), *causa.esposizione* indica la causa dell'esposizione ad amianto (per lavoro, per altre cause come per motivi ambientali o per hobby, o misto). *fibra.inalata* dice il tipo di fibra a cui il paziente è stato esposto (anfiboli, crisotilo o misto) mentre *livello.esposizione* classifica il livello di esposizione ad amianto (basso, medio, alto o altro);
- *anfiboli, crisotilo*: descrivono la percentuale di un determinato tipo di fibra trovato sul totale;
- *corpuscoli, corpuscoli.LL, corpuscoli.UL*: si riferiscono alla media geometrica del numero di corpuscoli all'interno dei campioni di polmone durante le analisi. È stata usata la media geometrica poichè la distribuzione del numero di corpuscoli non è normale. Le ultime due quantità sono gli estremi di un intervallo di confidenza al 95% di *corpuscoli* ottenuti dall'analisi effettuata tramite microscopio elettronico ottico per mezzo di un software dedicato. LL sta per "Lower Limit", ovvero "limite inferiore" ed UL sta per "Upper Limit", cioè "limite superiore";
- *fibre, fibre.LL, fibre.UL*: fanno riferimento ai risultati delle analisi di fibre nel tessuto polmonare. *fibre* è la media geometrica del numero di fibre, usata per lo stesso motivo di prima, su un grammo di tessuto secco di polmone, *fibre.LL e fibre.UL* sono gli estremi di un intervallo di confidenza al 95% di *fibre* ottenuti dall'analisi effettuata tramite microscopio elettronico ottico per mezzo di un software dedicato. Il significato di UL e LL è il medesimo del caso precedente.

2.2 Pulizia del dataset

Quando la matrice dei dati oggetto di studio presenta uno o più valori non osservati, si parla di problema di dati mancanti. Ad esempio chi compila un sondaggio può non sapere una risposta da inserire, durante un esperimento industriale ci possono essere dati mancanti a causa di un guasto di un'apparecchiatura, per un'indagine regionale possono non essere disponibili i dati di alcuni comuni.
Generalmente vengono distinti due tipi di dati mancanti [14]: i primi sono le mancate risposte totali ("MRT" o in inglese "unit nonresponse"), ovvero quando mancano tutti i dati relativi ad un certo soggetto. I secondi sono le mancate risposte parziali ("MRP" o in

inglese “item nonresponse”), cioè quando è presente solo qualche valore mancante in alcuni soggetti.
Esistono diversi metodi, di varia complessità, per risolvere il problema, a seconda della natura delle variabili e dal tipo di studio; per ovvi motivi solo alcuni di essi sono stati presentati ed utilizzati in questa tesi.

2.2.1 Dati mancanti nel dataset

Dal momento che il database originale presentava un numero di caratteristiche molto elevato è stato necessario operare una selezione delle informazioni più rilevanti. Un lavoro simile è stato effettuato per quanto riguarda le osservazioni, che inizialmente erano 534: sono stati eliminati infatti tutti i soggetti per cui non erano presenti le due misurazioni più importanti ovvero il numero di fibre e di corpuscoli. Non sono state considerate le osservazioni con problemi di codifica e che avevano come fonte Brescia, in quanto non erano casi investigati dal Registro Mesoteliomi del Veneto e le informazioni non erano sufficienti. È stato inoltre importante analizzare solo i dati provenienti dal laboratorio della dott.ssa Somigliana, in quanto è risaputo che per vari motivi i risultati provenienti da laboratori diversi possono essere non confrontabili.
Come si evince dalla Tabella 2.2.1 sono stati divisi i 534 soggetti in base ai quattro criteri principali ed è risultato che 341 soggetti presentavano almeno uno dei problemi (avevano come fonte Brescia, avevano problema di codifica o mancavano le misurazioni di fibre o corpuscoli). Di conseguenza le unità statistiche rimaste che rispettavano tutti i criteri di inclusione sono 193.

Criteri di esclusione	n	%
Problema fonte Brescia	139	26
Problema di codifica	1	0.18
Mancanza fibre	33	6.18
Mancanza corpuscoli	332	2.17
Soggetti iniziali	534	100
Soggetti esclusi	341	63.86
Soggetti analizzati	193	36.14

Tabella 2.2.1: Selezione dei soggetti tramite quattro criteri.

È iniziata successivamente la procedura di trattamento dei dati mancanti. Come si nota dalla Tabella 2.2.2 vi erano diverse caratteristiche che presentavano valori mancanti con valori massimi che raggiungevano il 38.9% di valori mancanti sul totale. È facile notare che non ci sono soggetti senza identificatore, senza numero di fibre o senza numero di corpuscoli, in quanto erano già stati eliminati nel passaggio precedente.

CARATTERISTICA	N	%	CARATTERISTICA	N	%
id	0	0	data.decesso	1	0.5
genere	1	0.5	settore.lavorativo	15	7.8
fonte	26	13.5	causa.esposizione	10	5.2
regione	32	16.6	fibra.inalata	75	38.9
malattia	12	6	livello.esposizione	47	24.4
intervento	15	8	anfiboli	10	5.2
stato.vita	0	0	crisotilo	10	5.2
data.nascita	11	5.7	corpuscoli	0	0
data.inizio	18	9.3	corpuscoli.LL	14	8.8
data.fine	17	8.8	corpuscoli.UL	14	8.3
data.prelievo	19	10	fibre	0	0
data.diagnosi	70	36.3	fibre.LL	17	7.3
data.analisi.corpuscoli	21	10.9	fibre.UL	16	7.3

Tabella 2.2.2: Numero di dati mancanti per ogni caratteristica e rispettiva percentuale sul totale.

È riportata una breve tabella riassuntiva dei metodi utilizzati a seconda del tipo di caratteristica; sono state usate diverse metodologie a seconda della loro natura e della disponibilità di informazioni nel dataset. Con "altre car." si intende che sono sono state stimate tramite altre caratteristiche, "regr." significa tramite un modello di regressione e "mice" per mezzo di MICE (che sta per "Multivariate Imputation by Chained Equations") che è un metodo di imputazione statistico. "no NA" sta per "nessun valore mancante".

CARATTERISTICA	METODO	CARATTERISTICA	METODO
id	no NA	data.decesso	altre car.
genere	altre car.	settore.lavorativo	altre car.
fonte	altre car.	causa.esposizione	altre car.
regione	altre car.	fibra.inalata	altre car.
malattia	altre car.	livello.esposizione	altre car.
intervento	altre car.	anfiboli	altre car.
stato.vita	no NA	crisotilo	altre car.
data.nascita	mice	corpuscoli	no NA
data.inizio	mice	corpuscoli.LL	regr.
data.fine	mice	corpuscoli.UL	regr.
data.prelievo	altre car.	fibre	no NA
data.diagnosi	mice	fibre.LL	regr.
data.analisi.corpuscoli	mice	fibre.UL	regr.

Tabella 2.2.3: Metodo di stima dei valori mancanti per ogni caratteristica.

2.2.2 Recupero tramite altre caratteristiche

Per una preliminare operazione di recupero dei dati mancanti sono state sfruttate dove possibile altre caratteristiche; il database di partenza contiene molti dati e quindi una grossa parte delle informazioni è stata recuperata da altre variabili o da fonti esterne.
Ad esempio per risalire alla regione di residenza è stato usato il comune di residenza, per stimare la causa d'esposizione è stato usato il settore lavorativo e la caratteristica

fibra.inalata è stata stimata sfruttando le percentuali di anfiboli e crisotilo. Allo stesso modo sono stati stimati alcuni valori mancanti di *genere*, *malattia*, *intervento* e *data.decesso*; dove non erano disponibili i valori di *data.prelievo* sono stati stimati con la data di analisi. Un procedimento analogo è stato condotto per le quantità temporali; quando infatti era disponibile solo l'anno di un certo evento è stata scelta arbitrariamente come data il 30 giugno di tale anno in quanto si tratta del giorno dell'anno mediano.
Dove non erano disponibili le percentuali di anfiboli e crisotilo trovate sono state stimate sfruttando la media di quelle disponibili.

2.2.3 Multivariate Imputation by Chained Equations (MICE)

Per stimare le date degli eventi (tranne *data.prelievo* e *data.decesso*) è stato usato l'approccio dell'imputazione multipla, che mira a ridurre l'incertezza sui dati mancanti creando diversi dataset imputati plausibili e combinando opportunamente i risultati ottenuti da ciascuno di essi. Dalla libreria di R "MICE", che come già detto sta per "Multivariate Imputation by Chained Equations", ovvero "imputazione multivariata per equazioni concatenate" è stata scelta per questa operazione la funzione *mice* [15], che genera algoritmi di imputazione per dati mancanti multivariati, sfruttando la correlazione con altre caratteristiche presenti all'interno del dataset.
Il processo delle equazioni concatenate segue una sequenza di procedure ben definita:

1. Come prima cosa, per ogni valore mancante all'interno del dataset viene eseguita un'imputazione semplice come ad esempio la media; tali valori fungono da "segnaposti".
2. I "segnaposti" di una sola variabile VAR1 vengono ripristinati come dati mancanti.
3. I valori osservati della variabile VAR1 sono regrediti sulle altre variabili del dataset, ovvero VAR1 gioca il ruolo di variabile risposta nel modello di regressione mentre le altre variabili rappresentano le esplicative.
4. I valori mancanti di VAR1 vengono imputati come previsioni del modello di regressione. Naturalmente quando VAR1 verrà in seguito utilizzata come variabile indipendente nei modelli di regressione per le altre variabili verranno sfruttati sia i valori osservati che quelli imputati.
5. I passaggi 2, 3 e 4 vengono ripetuti per ogni variabile del dataset che presenta valori mancanti. La procedura per ogni variabile costituisce un ciclo.
6. I passaggi 2, 3 e 4 vengono ripetuti per un numero di cicli, con le imputazioni aggiornate ad ogni ciclo.

Il numero di cicli dipende dal problema di imputazione; l'idea è che alla fine dei cicli la distribuzione dei parametri che governano le imputazioni (come i coefficienti nei modelli di regressione) dovrebbe convergere e diventare stabili. Ciò può evitare ad esempio la dipendenza dall'ordine in cui sono state imputate le variabili.

In seguito viene riportato un grafico[16] con le distribuzioni marginali delle 5 date calcolate grazie alla stima kernel di densità, divise in dati osservati (blu) e m=5 densità stimate (rosso). Come si può vedere, si notano alcuni picchi anomali in *data.inizio*, *data.fine* e *data.diagnosi*, più evidenti in quest'ultima caratteristica probabilmente perchè il numero di dati mancanti è maggiore.

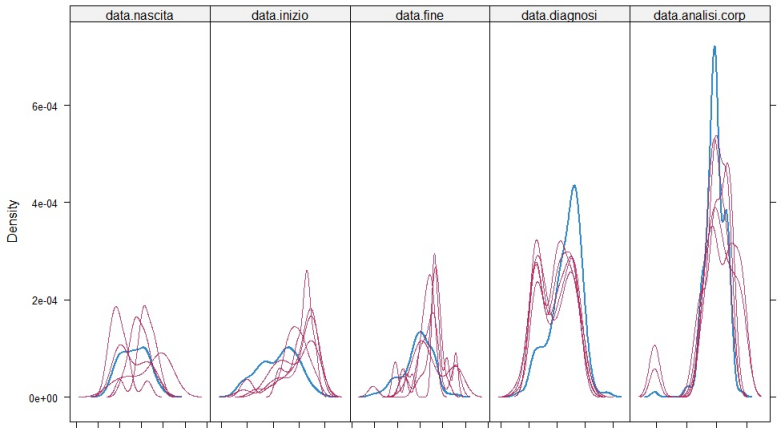

Figura 2.2.1: Stime kernel di densità per le distribuzioni marginali dei valori osservati (blu) e stimati (rosso).

Questa procedura ha i vantaggi di essere molto flessibile e può gestire variabili di varia natura (ad esempio continue e binarie). Uno svantaggio è che non ha la stessa giustificazione teorica di altri approcci di imputazione. In particolare, l'adattamento di una serie di distribuzioni condizionali, come ad esempio avviene utilizzando una serie di modelli di regressione, potrebbe non essere coerente con una corretta distribuzione congiunta.

2.2.4 Regressione

Gli intervalli di confidenza sulla stima del numero di fibre e corpuscoli ovvero *fibre.LL, fibre.UL, corpuscoli.LL* e *corpuscoli.UL* sono stati stimati tramite un modello di regressione [17], sfruttando il fatto che gli intervalli sono centrati sulle stime e che l'ampiezza dell'intervallo di confidenza è molto correlata con la quantità di fibre e di corpuscoli, come è evidente nella Figura 2.2.2. Di conseguenza i valori mancanti di *corpuscoli.LL* e *corpuscoli.UL* sono stati ottenuti tramite un modello di regressione lineare utilizzando come variabile dipendente il logaritmo dell'ampiezza degli intervalli di confidenza, vale a dire *log(corpuscoli.UL-corpuscoli.LL)* e come variabile dipendente il logaritmo della misurazione di corpuscoli, disponibile per tutti e 193 i soggetti. La scelta della trasformata logaritmica è dettata da una migliore interpretabilità grafica.
Il modello scelto inizialmente ed il corrispettivo output sono i seguenti:

$$\log(y_i) = \alpha + \beta \cdot \log(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

dove $y_i = ampiezza.corp_i = corpuscoli.UL_i - corpuscoli.LL_i$ ed $x_i = corpuscoli_i$.

	Dependent variable:
	log(y)
log(x)	0.777***
	(0.025)
Constant	2.234***
	(0.261)
Observations	177
R^2	0.847
Adjusted R^2	0.846
Akaike Inf. Crit.	390.064
Bayesian Inf. Crit.	399.593
Residual Std. Error	0.720 (df = 175)
F Statistic	970.848*** (df = 1; 175)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabella 2.2.4: Tabella riassuntiva del modello sui corpuscoli.

Per valutare la bontà dell'adattamento del modello è stato osservato il coefficiente di determinazione R^2, che varia da 0 a 1 ed è calcolato come

$$R^2 = 1 - \frac{SQ_{res}}{SQ_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

dove SQ_{res} ed SQ_{tot} sono rispettivamente la devianza residua e quella totale. Naturalmente, il modello si adatta bene se i valori stimati $\hat{y}_i$ si avvicinano a quelli reali y_i, ovvero quando SQ_{res} è piccola e quindi R^2 è vicino ad 1. In questo caso l'indice è pari a 0.847, quindi si può concludere che il modello si adatta bene ai dati.
Si nota inoltre che il modello presenta tutte le esplicative significative. Le stime dei parametri ottenute sono le seguenti:

$$\hat{\alpha} = 2.23418 \quad e \quad \hat{\beta} = 0.7772$$

Il modello stimato risulta essere quindi

$$\log(\hat{y}_i) = \hat{\alpha} + \hat{\beta} \cdot \log(x_i)$$

e di conseguenza le stime dell'ampiezza degli intervalli:

$$\hat{ampiezza.corp_i} = \hat{y}_i = e^{\hat{\alpha}+\hat{\beta}\cdot\log(x_i)}$$

Infine, sapendo che gli intervalli sono centrati sul valore di *corpuscoli*, i valori degli estremi sono stati calcolati tramite:

$$\hat{corpuscoli.LL_i} = corpuscoli_i - \frac{\hat{ampiezza.corp_i}}{2}$$

$$\hat{corpuscoli.UL_i} = corpuscoli_i + \frac{\hat{ampiezza.corp_i}}{2}$$

Similarmente i valori mancanti di *fibre.LL* e *fibre.UL* sono stati ottenuti tramite un modello di regressione quadratico utilizzando come variabile dipendente il logaritmo dell'ampiezza degli intervalli di confidenza ovvero *log(fibre.UL-fibre.LL)* e come variabile dipendente il logaritmo della misurazione di fibre.
In basso è riportato il modello scelto ed il rispettivo output.

$$\log(y_i) = \alpha + \beta \cdot \log(x_i) + \gamma \cdot (\log(x_i))^2 + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

dove $y_i = ampiezza.fibre_i = fibre.UL_i - fibre.LL_i$ ed $x_i = corpuscoli_i$.

	Dependent variable:
	log(y)
log(x)	-0.505^{***}
	(0.165)
I(log(x)^2)	0.042^{***}
	(0.005)
Constant	13.127^{***}
	(1.234)
Observations	179
R^2	0.963
Adjusted R^2	0.963
Akaike Inf. Crit.	16.688
Bayesian Inf. Crit.	29.437
Residual Std. Error	0.250 (df = 176)
F Statistic	$2{,}315.636^{***}$ (df = 2; 176)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabella 2.2.5: Tabella riassuntiva del modello sulle fibre.

Anche in questo caso tutte le esplicative risultano significative ed il modello presenta un indice di determinazione R^2 alto (0.963). Si è inoltre scelto il modello quadratico piuttosto che il modello lineare dopo l'esame di alcuni indici.
L'indice di determinazione R^2 valuta la bontà di adattamento di un singolo modello e non è adatto per il confronto tra modelli in quanto aumenta all'aumentare del numero di esplicative, a prescindere dalla loro significatività. Per questo motivo è stato sfruttato il coefficiente di determinazione corretto $\overline{R^2}$, che oltre alla bontà di adattamento tiene conto anche del principio di parsimonia, ovvero tende a non scegliere modelli con un numero eccessivamente alto di esplicative.
Chiamato M_k un modello con k termini esclusa l'intercetta, quindi con $k-1$ esplicative, viene supposto che la stima del modello fornisca la somma dei quadrati dei residui $SQ_{res}^{M_k}$. L'indice è calcolato con la formula sottostante e massimizzarlo significa preferire il modello che minimizza s^2, la stima corretta di σ^2, tra tutti i modelli considerati.

$$\overline{R^2} = \frac{SQ_{res}^{M_k}/(n-k)}{SQ_{tot}/(n-1)}$$

Altri indici presi in considerazione sono l'AIC (Akaike Information Criterion, "Criterio di informazione di Akaike") ed il BIC (Bayesian Information Criterion, "Criterio di informazione bayesiano"), calcolati come $AIC = -2\hat{l}_{M_k} + 2k$ e $BIC = -2\hat{l}_{M_k} + k \cdot log(n)$, dove $\hat{l}_{M_k}$ è la log-verosimiglianza massimizzata per il modello M_k con k termini. Come si può facilmente vedere sono entrambi basati sulla verosimiglianza, rendendoli facilmente applicabili anche a modelli diversi da quelli lineari, e contengono una penalità per poter applicare il principio di parsimonia. Un modello è preferibile ad un altro quando il suo AIC o BIC è minore.
Il modello di regressione quadratico è stato preferito al modello di regressione lineare in quanto presenta un coefficiente di determinazione corretto $\overline{R^2}$ maggiore (0.963 contro 0.951) un AIC minore (16.7 contro 66) ed un BIC minore (29.4 contro 75.5).
Inoltre, per valutare o meno l'inclusione del termine quadratico è stato usato un test ANOVA per il confronto di modelli annidati.
Si tratta di un test parametrico per il confronto di due modelli annidati (uno compreso nell'altro), in generale tra il modello saturo $M_1 : Y = XB + \varepsilon$ che contiene p parametri ed il modello ridotto M_0 che ne contiene $p_0 < p$.
Verifica il sistema d'ipotesi sulla nullità di un gruppo di $p - p_0$ coefficienti, ovvero:

$$\begin{cases} H_0 : \beta_{p_0+1} = \beta_{p_0+2} = ... = \beta_p = 0 \\ H_1 : \exists! r \in (p_{0+1}, ..., p) | \beta_r \neq 0 \end{cases}$$

Più nello specifico si ha che il modello saturo è

$$Y \sim N_n(X\beta, \sigma^2 I) \qquad Y = X\beta + \varepsilon \qquad \hat{\sigma}^2 = \frac{e^T e}{n}$$

dove Y è il vettore delle esplicative di lunghezza n, n è il numero di osservazioni, X è la matrice dei regressori di dimensione $n \times p$, β è il vettore di parametri di lunghezza p, $\sigma^2 I$ è la matrice di varianze e covarianze di dimensione $n \times n$ del vettore degli errori ε di lunghezza n. e è il vettore dei residui di lunghezza n mentre $\hat{\sigma}^2$ è una stima di σ^2. Il modello ridotto è invece

$$Y \sim N_n(X_0\beta_0, \sigma^2 I) \qquad Y = X_0\beta_0 + \varepsilon_0 \qquad \hat{\sigma}_0^2 = \frac{e_0^T e_0}{n}$$

dove Y è il vettore delle esplicative di lunghezza n, n è il numero di osservazioni, X è la matrice dei regressori di dimensione $n \times p_0$, β è il vettore di parametri di lunghezza p_0, $\sigma^2 I$ è la matrice di varianze e covarianze di dimensione $n \times n$ del vettore degli errori ε_0 di lunghezza n. e_0 è il vettore dei residui di lunghezza n mentre $\hat{\sigma}_0^2$ è una stima di σ^2.
Dal momento che si vuole quantificare la perdita di bontà di adattamento del modello M_0 rispetto ad M_1, la statistica test sarà

$$F = \frac{(SQ_{res}^0 - SQ_{res}^1)/(p - p_0)}{(SQ_{res}^1)/(n-p)} = \frac{(e_0^T e_0 - e_1^T e_1)/(p-p_0)}{e_1^T e_1/(n-p)} = \frac{(\hat{\sigma}_0^2 - \hat{\sigma}_1^2)/(p-p_0)}{\hat{\sigma}_1^2/(n-p)}$$

che sotto H_0 si distribuisce come una $F_{p-p_0,n-p}$. Con un p-value molto piccolo si rifiuta l'ipotesi nulla di nullità del termine quadratico.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
177	14.65	NA	NA	NA	NA
176	11	1	3.65	58.39	1.329e-12

Tabella 2.2.6: ANOVA per modelli annidati per valutare la significatività del termine quadratico.

Le stime dei parametri ottenute dal modello sono le seguenti:

$$\hat{\alpha} = 13.127468, \quad \hat{\beta} = -0.504868, \quad \hat{\gamma} = 0.041884$$

Il modello stimato è dunque

$$\log(\hat{y}_i) = \hat{\alpha} + \hat{\beta} \cdot \log(x_i) + \hat{\gamma} \cdot (log(x_i))^2$$

Ne consegue chiaramente che

$$\hat{ampiezza.fibre_i} = \hat{y}_i = e^{\hat{\alpha}+\hat{\beta}\cdot\log(x_i)+\hat{\gamma}\cdot(log(x_i))^2}$$

Infine, le stime degli estremi degli intervalli sono

$$\hat{fibre.LL_i} = fibre_i - \frac{\hat{ampiezza.fibre_i}}{2}$$

$$\hat{fibre.UL_i} = fibre_i + \frac{\hat{ampiezza.fibre_i}}{2}$$

Nelle figure sottostanti sono rappresentati i valori osservati ed i valori stimati dai rispettivi modelli.

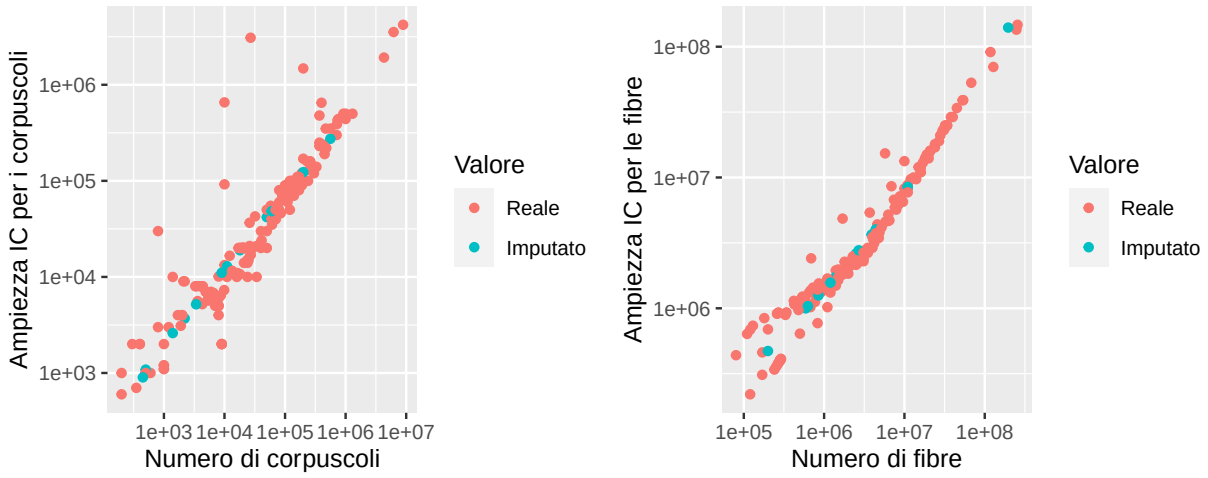

Figura 2.2.2: Relazioni tra misurazioni (di corpuscoli e fibre) e ampiezza dei rispettivi intervalli di confidenza (scala logaritmica).

Il metodo della regressione (o Buck's method) [18] è piuttosto comodo ma non senza difetti. In primo luogo la relazione tra l'esplicativa e la risposta deve essere molto alta affinchè il modello restituisca delle buone stime; in questo caso, come si è visto, questa assunzione era garantita. Inoltre richiede l'assunzione di normalità delle variabili (sono stati utilizzati due modelli di regressione normali) ed in questo caso era garantita. Inoltre questo metodo è sensibile ai dati anomali, non è esente da distorsioni e si può correre il rischio che vengano imputati valori non reali rispetto al problema preso in considerazione.

2.3 Test statistici

In questo paragrafo vengono presentati i test statistici utilizzati per le successive analisi.

- Il test di Shapiro-Wilk è considerato uno dei test più potenti per verificare la normalità di un campione di dati, anche se di piccole dimensioni, e presenta come sistema di ipotesi
$$\begin{cases} H_0 : I \ dati \ sono \ distribuiti \ normalmente \\ H_1 : I \ dati \ non \ sono \ distribuiti \ normalmente \end{cases}$$
La statistica test su cui si basa è:
$$W_{SW} = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
dove n è il numero di osservazioni, x_i sono le osservazioni, $x_{(i)}$ sono le osservazioni ordinate, $\overline{x}$ è la media aritmetica del campione ed a_i sono delle costanti disponibili su apposite tabelle. La statistica W_{SW} può essere interpretata come il quadrato del coefficiente di correlazione del grafico quantile-quantile delle osservazioni, di conseguenza è una misura della relazione lineare. Infatti si rifiuta l'ipotesi nulla di normalità se il valore della statistica test si discosta significativamente da 1.

- Il test utilizzato per verificare l'omogeneità delle varianze in J diversi gruppi è il test di Levene, che ha come sistema di ipotesi:
$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_J^2 \\ H_1 : \exists \, ! \, \sigma_h^2 \neq \sigma_k^2 \, \forall \, (\sigma_h^2, \sigma_k^2), \quad h \neq k \end{cases}$$
La statistica test è la seguente:
$$t_L = \frac{(n-J)\sum_{j=1}^{J} n_j(\overline{z_j} - \overline{z})^2}{(J-1)\sum_{j=1}^{J}\sum_{i=1}^{n_j}(z_{ij} - \overline{z_j})^2}$$
dove n è il numero di osservazioni, J è il numero di gruppi, n_j è la numerosità del gruppo j-esimo, $\overline{z}$ è la media complessiva, $\overline{z_j}$ è pari a $\sum_{i=1}^{n_j} \frac{z_{ij}}{n_j}$ e z_{ij} può assumere i seguenti valori (x_{ij} sono le osservazioni):

1. z_{ij}=$|x_{ij} - \overline{x_j}|$ con $\overline{x_j}$ media del gruppo j-esimo;
2. z_{ij}=$|x_{ij} - m_{0.5,_j}|$ con $m_{0.5,_j}$ mediana del gruppo j-esimo.
3. z_{ij}=$|x_{ij} - \overline{x_j^t}|$ con $\overline{x_j^t}$ media troncata del gruppo j-esimo. La media troncata si ottiene togliendo una piccola percentuale fissata di valori minimi e/o massimi, ovvero i valori considerati anomali.

Sotto H_0 la statistica test si distribuisce come una $F_{J-1,n-J}$.

Per verificare la differenza in media del logaritmo del numero di corpuscoli nei vari gruppi di esplicative sono stati usati due test: l'ANOVA a una via (parametrico) e l'ANOVA per ranghi a una via di Kruskal-Wallis [19] (non parametrico).

- Il primo è un test che permette di confrontare le medie di tre o più gruppi. Siano J il numero di gruppi da confrontare: si assume che le y_{ij} siano realizzazioni di variabili casuali normali indipendenti, ovvero $Y \sim N(\mu_j, \sigma^2)$ con $j = 1, ..., J$. L'obiettivo di interesse è verificare l'ipotesi nulla di uguaglianza delle J medie:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_J \\ H_1 : \exists!(\mu_h, \mu_k) | \mu_h \neq \mu_k, \quad h \neq k \end{cases}$$

  Il test ANOVA assume quindi la normalità della risposta nei diversi gruppi e l'omogeneità delle varianze della risposta nei gruppi.
  La statistica test utilizzata è:

$$F = \frac{\sum_{j=1}^{J} n_j(\overline{y_j} - \overline{y})^2/(J-1)}{\sum_{j=1}^{J}\sum_{i=1}^{n}(y_{ij} - \overline{y_j})^2/(n-J)}$$

  dove n è il numero di osservazioni, J è il numero di gruppi, n_j è la numerosità del gruppo j-esimo, y_{ij} sono le osservazioni, $\overline{y}$ è la media generale e $\overline{y_j}$ è la media del gruppo j-esimo. Sotto H_0 la statistica test si distribuisce come una $F_{J-1,n-J}$.

- L'ANOVA per ranghi a una via di Kruskal-Wallis è un test non parametrico basato sui ranghi per verificare l'uguaglianza delle mediane in J diversi gruppi; non richiede alcuna assunzione distributiva. Il sistema d'ipotesi è il seguente:

$$\begin{cases} H_0 : m_{0.5,\,1} = m_{0.5,\,2} = ... = m_{0.5,\,J} \\ H_1 : \exists!(m_{0.5,\,h}, m_{0.5,\,k}) | m_{0.5,\,h} \neq m_{0.5,\,k}, \quad h \neq k \end{cases}$$

  Il rango r_i è il numero intero che identifica la posizione che un dato x_i occupa se si ordina il campione casuale in senso crescente, ovvero passando da $(x_1, ..., x_n)$ a $(x_{(1)}, ..., x_{(n)})$. Il procedimento per applicare questo test è il seguente, supponendo di avere n osservazioni chiamate x_{ij} e J gruppi.
  Le osservazioni x_{ij} vengono ordinate in senso crescente e vengono assegnati i ranghi da 1 ad n; i ranghi assegnati ai valori in ognuno degli J gruppi vengono sommati tra di loro, ottenendo J somme di ranghi chiamate R_j, $j = 1, ..., J$. Infine per ogni gruppo

sono calcolati i ranghi medi $\overline{R_j}$, $j = 1, ..., J$ e la media generale $\overline{R} = \frac{n+1}{2}$.
La statistica test è calcolata come:

$$t_{KW} = \frac{12}{n(n+1)} \sum_{j=1}^{J} n_j (\overline{R_j} - \overline{R})^2 = \frac{12}{n(n+1)} \sum_{j=1}^{J} \frac{R_j^2}{n_j} - 3(n+1)$$

Se sono presenti solo tre gruppi ed ogni gruppo contiene cinque o meno dati, la significatività della statistica test t_{KW} è determinata da un'opportuna tabella. Se invece ogni gruppo contiene più di cinque osservazioni, sotto H_0 la statistica test t_{KW} si distribuisce come una χ^2_{J-1}.

Una volta eseguiti i test per la verifica di uguaglianza del logaritmo del numero di corpuscoli nei vari gruppi, i test post-hoc (o test a posteriori) permettono di individuare le coppie di gruppi statisticamente differenti tra loro. I test utilizzati in questo caso sono due: il metodo di Holm per l'ANOVA a una via ed il test di Mann-Whitney per ogni coppia per l'ANOVA per ranghi a una via di Kruskal-Wallis.

- Il metodo di Holm consiste in una procedura di rifiuto sequenziale ed ha lo scopo di correggere il p-value in caso di confronti multipli. Chiamato J il numero di gruppi, vengono svolte tutte le $\frac{J(J-1)}{2}$ verifiche d'ipotesi a coppie tramite il test t, calcolando i rispettivi p-value. Successivamente vengono ordinati in senso crescente, ottenendo $p - value_{(1)} \leq ... \leq p - value_{(j)} \leq p - value_{\frac{J(J-1)}{2}}$ ed il primo p-value viene corretto attraverso la procedura di Bonferroni, ossia viene moltiplicato per il numero di confronti effettuati. Si rifiuta poi se $p - value_{(1)} \leq \alpha_1$ ed il j-esimo p-value $p - value_{(j)}$ viene confrontato con $\frac{\alpha}{\frac{J(J-1)}{2} - j + 1}$, rifiutando se $p - value_{(j)} < \frac{\alpha}{\frac{J(J-1)}{2} - j + 1}$. Si procede finchè non si rifiuta H_0 oppure si esauriscono le ipotesi, cioè quando tutti i confronti a coppie sono significativi.

- In caso di rifiuto dell'ipotesi nulla nell'ANOVA per ranghi a una via di Kruskal-Wallis, per capire quali sono le coppie statisticamente differenti tra loro si può procedere con il test di Mann-Whitney per ogni coppia, correggendo i p-value con il metodo di Holm. Il test di Mann-Whitney è un test non parametrico che, dati due campioni casuali semplici $x_1 = (x_{11}, ..., x_{1n_1})$ e $x_2 = (x_{21}, ..., x_{2n_2})$ di lunghezza n_1 ed n_2, verifica il sistema di ipotesi

$$\begin{cases} H_0 : m_{0.5}(x_1) = m_{0.5}(x_2) \\ H_1 : m_{0.5}(x_1) \neq m_{0.5}(x_2) \end{cases}$$

o equivalentemente

$$\begin{cases} H_0 : F_{X_1}(x) = F_{X_2}(x) \\ H_1 : F_{X_1}(x) \neq F_{X_2}(x) \end{cases}$$

dove $F(x)$ è la funzione di ripartizione.
La statistica test è $t_{MW} = S_1 - \frac{n_1(n_1+1)}{2}$ con $S1 = \sum_{i=1}^{n_1} r_{1i}$; r_{1i} è il rango che compete ad x_{1i} nel campione combinato ed ordinato ottenuto ordinando in senso crescente i valori di x_1 ed x_2. Una volta calcolato il valore di t_{MW}^{OSS} si consultano le tavole dei valori

critici per vari valori di n_1 ed n_2. Se n_1 ed n_2 sono sufficientemente grandi, si può sfruttare il Teorema del Limite Centrale ed utilizzare la statistica test

$$t^*_{MW} = \frac{t_{MW} - \frac{n_1 n_2}{2} - 0.5}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

che sotto H_0 segue approssimativamente una distribuzione $N \sim (0, 1)$.

2.4 Analisi univariate

In questo paragrafo viene effettuata un'analisi descrittiva univariata delle caratteristiche a disposizione tramite tabelle e grafici esplicativi per dare un'idea preliminare degli attributi dei soggetti. Nelle due tabelle sottostanti si vede una netta prevalenza degli uomini (94%) rispetto alle donne (6%); inoltre solo 18 soggetti su 193 sono vivi al momento della raccolta dei dati e questo fatto spiega perchè la maggior parte dei campioni di tessuto polmonare sono stati rimossi tramite autopsia (155 su 193). Le patologie più frequenti sono il mesotelioma ed il tumore al polmone, in linea con la letteratura scientifica per quanto riguarda le malattie asbesto correlate. Le regioni di residenza più frequenti (Friuli e Veneto) si riflettono sulla fonte dei dati.

	Stato.vita			Genere			Intervento			Malattia	
	N	%		N	%		N	%		N	%
Vivi	18	9	Uomo	181	94	Biopsia	38	20	Mesotelioma	127	65.8
Deceduti	175	91	Donna	12	6	Autopsia	155	80	Tumore polmonare	55	28.5
									Altro	11	5.7

Tabella 2.4.1: Caratteristiche preliminari e mediche dei soggetti.

	Fonte			Regione	
	N	%		N	%
Gorizia	105	54.4	Friuli	101	52.3
Padova	55	28.5	Veneto	55	28.5
Altro	33	17.1	Altro	37	19.2

Tabella 2.4.2: Caratteristiche demografiche dei soggetti.

Nella Tabella 2.4.3 vengono riportate delle statistiche di sintesi della storia dei soggetti. La maggior parte di loro è nata tra il 1930 ed il 1944 ed ha cominciato l'esposizione circa 20 anni dopo, tra il 1951 ed il 1969. La fine dell'esposizione avviene prevalentemente tra il 1978 ed il 1990, poco prima che in Italia venisse vietata la produzione di amianto. Gli anni del prelievo, della diagnosi, dell'analisi del tessuto polmonare e del decesso sono piuttosto simili e si aggirano nei primi due decenni del 2000. È importante ricordare che 18 soggetti sono vivi al momento della raccolta dei dati e che quindi per loro non viene riportata la data del decesso. Analoghe conclusioni si possono dedurre dalla Tabella 2.4.4: i soggetti cominciano ad essere esposti in media a circa 23 anni e l'esposizione termina dopo circa 20

anni. Le età in cui avvengono prelievo, diagnosi, analisi e decesso si concentrano tra i 64 e gli 82 anni. Il periodo di latenza, ovvero la distanza temporale che intercorre tra l'anno medio dell'esposizione ed il prelievo, ha una mediana di 39 anni, con un minimo di 20.5 anni ed un massimo di 58, mentre la durata dell'esposizione è prevalentemente compresa tra 13 e 32 anni.
La tabella, così come le successive, riporta il valore minimo, il primo quartile, la mediana, la media, il terzo quartile, il valore massimo e la deviazione standard.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Anno di nascita	1918	1930	1938	1938	1944	1959	8.7
Anno di inizio esp.	1937	1951	1962	1960	1969	1980	10.91
Anno di fine esp.	1954	1978	1984	1983	1990	2005	9.8
Anno di prelievo	2002	2010	2011	2011	2012	2018	2.5
Anno di diagnosi	2001	2008	2010	2009	2011	2018	3.2
Anno di analisi	2002	2012	2013	2013	2015	2018	2.03
Anno decesso	2002	2009	2011	2011	2012	2018	2.66

Tabella 2.4.3: Statistiche di sintesi della storia dei soggetti.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Età inizio	10	16	20	22.62	27	52	8.41
Età fine	18	36	48	45.25	54	83	11.94
Età prelievo	51	67	73	73.15	80	93	8.77
Età diagnosi	51	64	71	71.46	78	93	9.18
Età analisi	51	69	75	75.48	82	95	8.92
Età decesso	51	67	73.5	73.55	80	93	8.51
Durata esp.	1	13	23	22.63	32	46	11.7
Periodo latenza	20.5	33	39	39.21	45.5	58	8.58

Tabella 2.4.4: Statistiche di sintesi degli eventi dei soggetti (valori espressi in anni).

La seguente tabella descrive le caratteristiche dell'esposizione ad amianto dei soggetti. Gli ambienti lavorativi più diffusi sono quelli dei cantieri navali, dove l'amianto è stato largamente usato. Circa l'87% delle unità statistiche è entrata in contatto con amianto per lavoro ed il livello di esposizione più diffuso è medio. Le fibre di anfiboli sono più diffuse del crisotilo, ma il 70% dei soggetti presenta un tipo di fibra misto.

	esp.sett			causa.esposizione			fibra.inalata			livello.esposizione	
	N	%		N	%		N	%		N	%
Cantieri navali	115	59.6	Lavoro	167	86.5	Anfiboli	51	26	Basso	34	17.6
Mezzi ferroviari	16	8.3	Altro	21	10.9	Crisotilo	8	4	Medio	80	41.4
Impianti industriali	11	5.7	Misto	5	2.6	Misto	134	70	Alto	32	16.6
Altro	51	26.4							Altro	47	24.4

Tabella 2.4.5: Caratteristiche dell'esposizione ad amianto dei soggetti.

La tabella sottostante mostra che in solo 3 soggetti non sono state trovate tracce di anfiboli, mentre circa la metà dei campioni di tessuto polmonare non presentavano fibre di crisotilo.

Tipo di fibra trovato	N	%
Anfiboli	190	98
Crisotilo	92	47

Tabella 2.4.6: Presenza del tipo di fibre.

Dalla Tabella 2.4.7 si può facilmente notare il maggior numero di fibre trovato rispetto ai corpuscoli, questo perchè non tutte le fibre subiscono il processo che le trasformano in corpuscoli di amianto. Si ottiene una conclusione analoga per l'ampiezza degli intervalli di confidenza, ovvero *Ampiezza.Corp* ed *Ampiezza.Fibre*, calcolati rispettivamente con *Ampiezza.Corp=corpuscoli.UL-corpuscoli.LL* ed *Ampiezza.Fibre=fibre.UL-fibre.LL*. È piuttosto alta la variabilità delle stime ed essendo le quantità *fibre* e *corpuscoli* dei conteggi che seguono quindi una variabile casuale di Poisson ci si aspetta una vicinanza tra media e varianza delle stime, che per una prima analisi non sembra esserci.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Corpuscoli	200	7000	26000	212678	130000	8900000	847039.5
Corpuscoli.LL	0	3000	18000	156687	1e+05	6790000	646432.8
Corpuscoli.UL	700	10000	40000	302134	190000	1.1e+07	1089787
Ampiezza.Corp	600	7000	20000	145447	90000	4210000	486180
Fibre	80000	840000	2200000	11685003	9e+06	2.56e+08	32430268
Fibre.LL	0	3e+05	1240000	7903195	6100000	1.9e+08	23394811
Fibre.UL	230000	1720000	3600000	16517456	1.3e+07	3.37e+08	43201927
Ampiezza.Fibre	220000	1370000	2250000	8614262	6950000	1.47e+08	20053782

Tabella 2.4.7: Statistiche di sintesi delle misurazioni in laboratorio di corpuscoli e fibre di amianto.

Il campione contiene 181 soggetti uomini (il 94%) e 12 donne (il 6%), di cui al momento della raccolta dati risultano 18 vivi (il 9%) e 175 deceduti (il 91%). Risiedono principalmente in Friuli Venezia Giulia e Veneto e similarmente la fonte dei dati è rappresentata quasi completamente dal tribunale di Gorizia e dal Registro Regionale Veneto dei casi di mesotelioma, con sede a Padova. Altri casi provengono da Trieste o Pistoia.

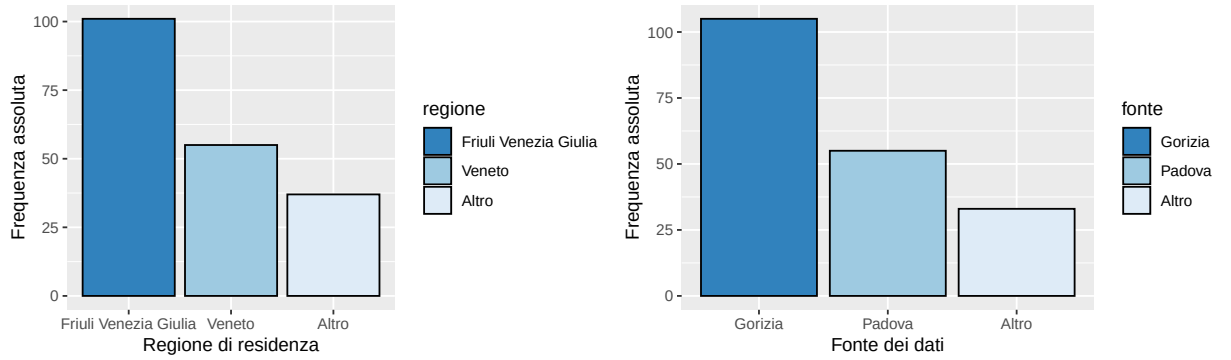

Figura 2.4.1: Distribuzioni della residenza e della fonte dei dati.

Come è ampiamente prevedibile, le patologie più frequenti sono il mesotelioma ed il tumore al polmone, mentre come già ribadito le autopsie prevalgono sulle biopsie.

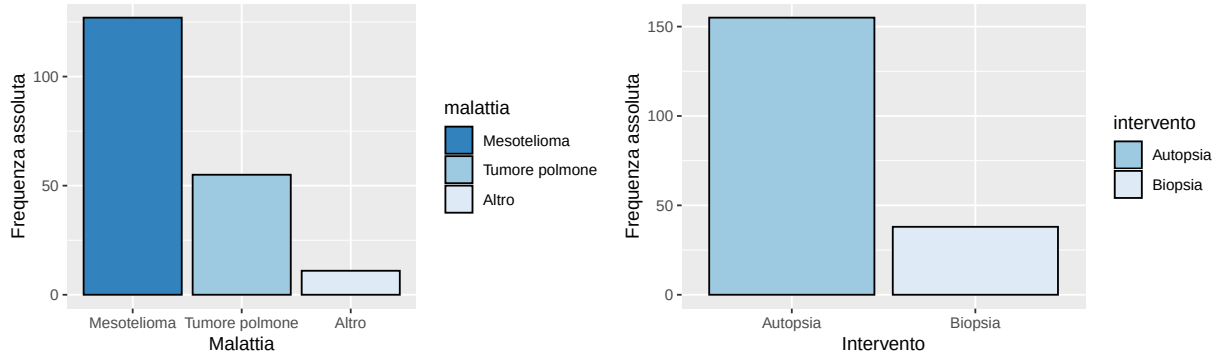

Figura 2.4.2: Distribuzioni della malattia e del tipo di intervento.

Come si può notare, la data di nascita si concentra negli anni attorno agli anni '30, mentre la data di inizio dell'esposizione si trova in gran parte negli anni '50 e '60, quando ci fu un aumento della produzione di amianto in Italia. Quasi tutti i soggetti terminano l'esposizione ad amianto prima del 1992, quando vennero vietate la produzione e l'installazione di materiali contenenti amianto.

I soggetti cominciano ad essere esposti ad amianto attorno ai 20 anni e terminano dopo circa 20 anni, principalmente tra i 36 ed i 54 anni.

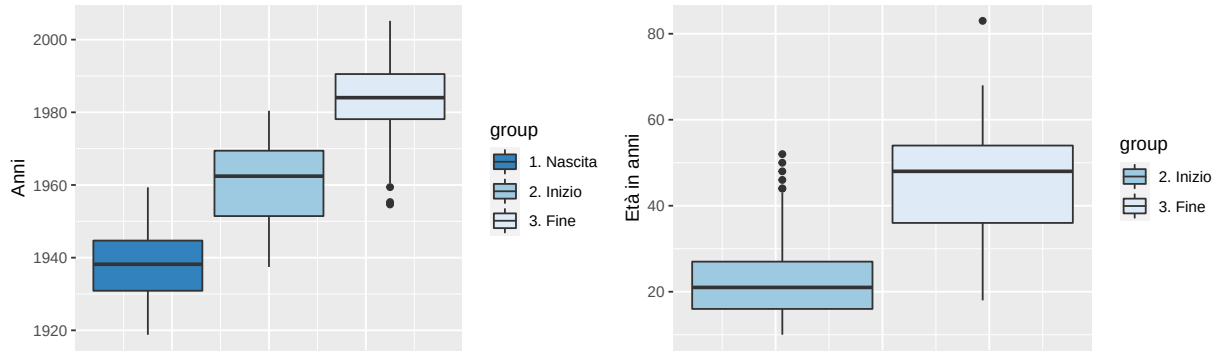

Figura 2.4.3: Distribuzioni della storia lavorativa dei soggetti.

Le date di prelievo, diagnosi, analisi e decesso sono situate nei primi due decenni del 2000, principalmente tra il 2008 ed il 2015 e si vede che in media l'analisi del tessuto polmonare avviene dopo della morte, questo perchè in 155 soggetti su 193 è stata effettuata l'asportazione del campione di polmone tramite autopsia, solo in 38 tramite biopsia. È facile inoltre osservare che per ovvi motivi il prelievo del tessuto polmonare avviene prima dell'analisi.
Le età dei quattro eventi sembrano essere piuttosto simili in media ed in varianza; variano per la maggior parte tra i 64 e gli 82 anni.

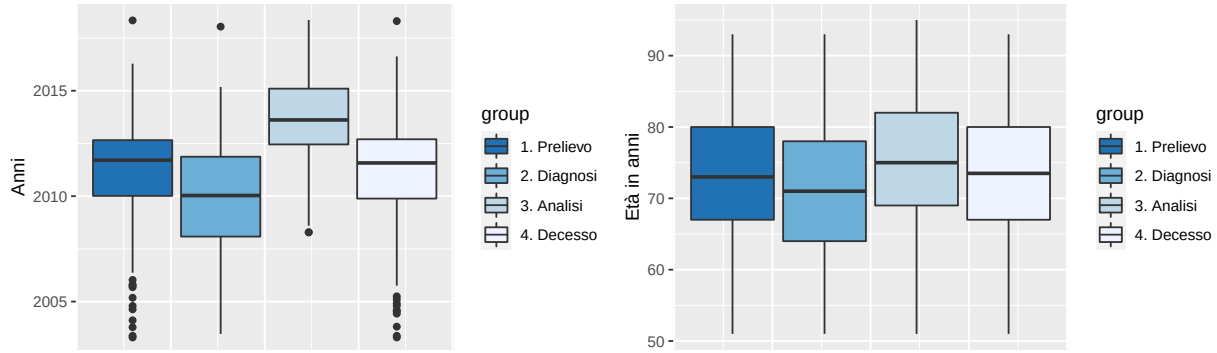

Figura 2.4.4: Distribuzioni della storia medica dei soggetti.

La durata di esposizione va un minimo di 1 anno ad un massimo di 46, ma si concentra tra i 12 ed i 32 anni; il periodo di latenza dura principalmente tra i 33 ed i 45 anni.

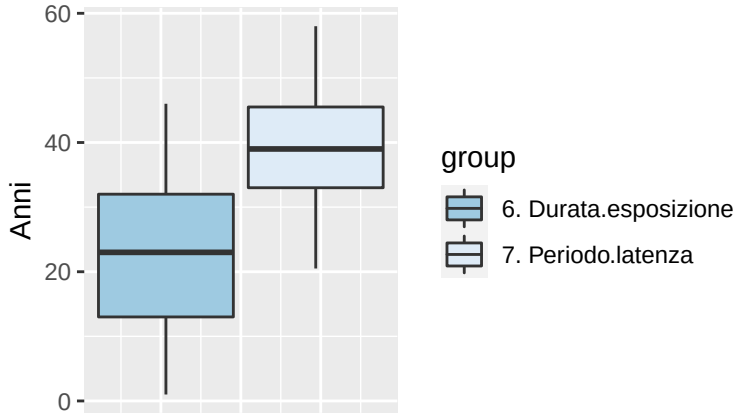

Figura 2.4.5: Distribuzione della durata dell'esposizione e del periodo di latenza.

Analizzando le caratteristiche relative all'esposizione ad amianto, si può notare una netta prevalenza di persone che sono entrate a contatto con l'amianto per lavoro, in particolar modo nei cantieri navali. Sono state trovate più fibre di anfiboli che di crisotilo, anche se più della metà delle osservazioni presenta un tipo di fibre misto, mentre il livello di esposizione più frequente è medio.

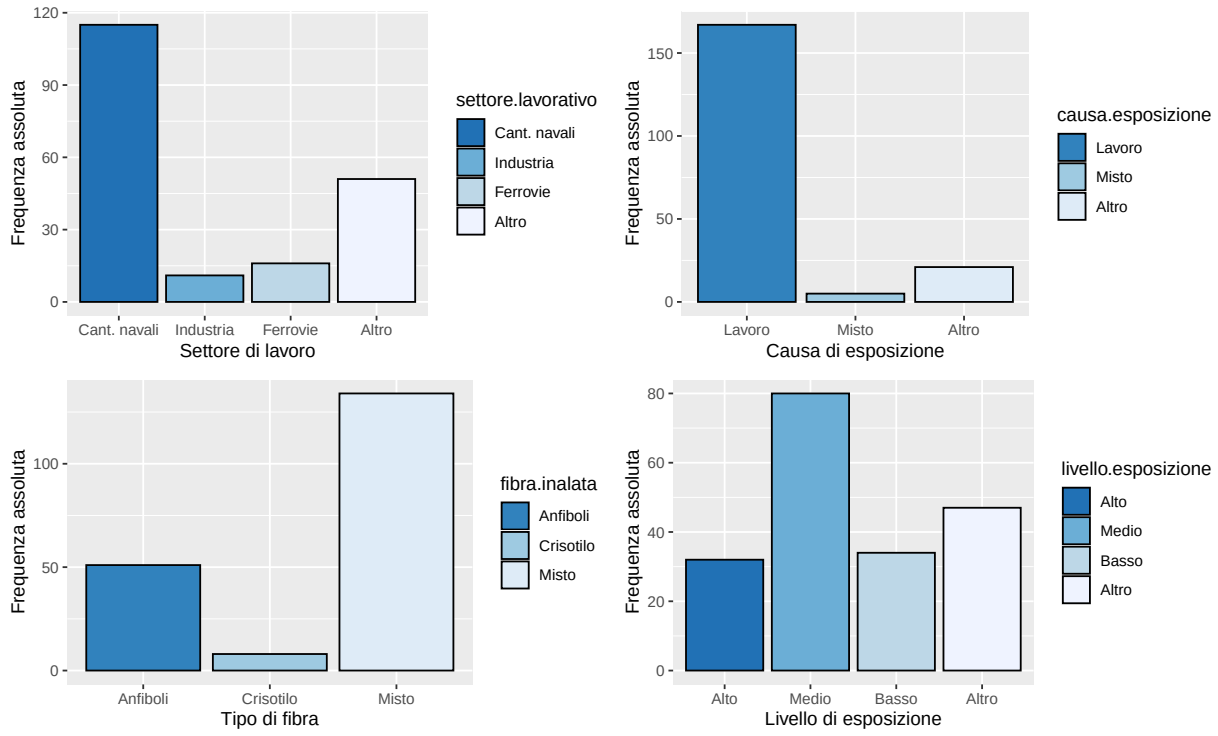

Figura 2.4.6: Distribuzioni del settore lavorativo, della causa dell'esposizione, del tipo di fibra e del livello di esposizione.

Riguardo il tipo di fibra rilevato, si nota che gli anfiboli sono molto frequenti, osservati nel 98.4% dei soggetti, mentre il crisotilo è meno frequente, trovato nel 47.66% di loro. Il 92% delle osservazioni presenta più anfiboli che crisotilo.

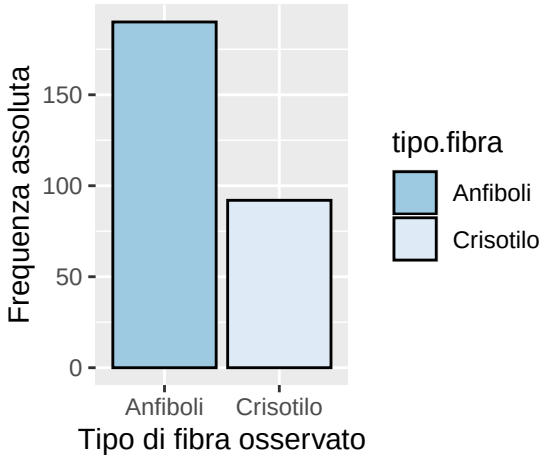

Figura 2.4.7: Diagramma a barre del tipo di fibra trovato.

Come si può facilmente notare dai grafici sottostanti, le distribuzioni del numero di corpuscoli e di fibre sono fortemente asimmetriche a destra. Per questo motivo si è optato per una trasformata logaritmica dei due conteggi, ottenendo una maggiore interpretabilità dei grafici

e delle proprietà distributive di normalità molto importanti.

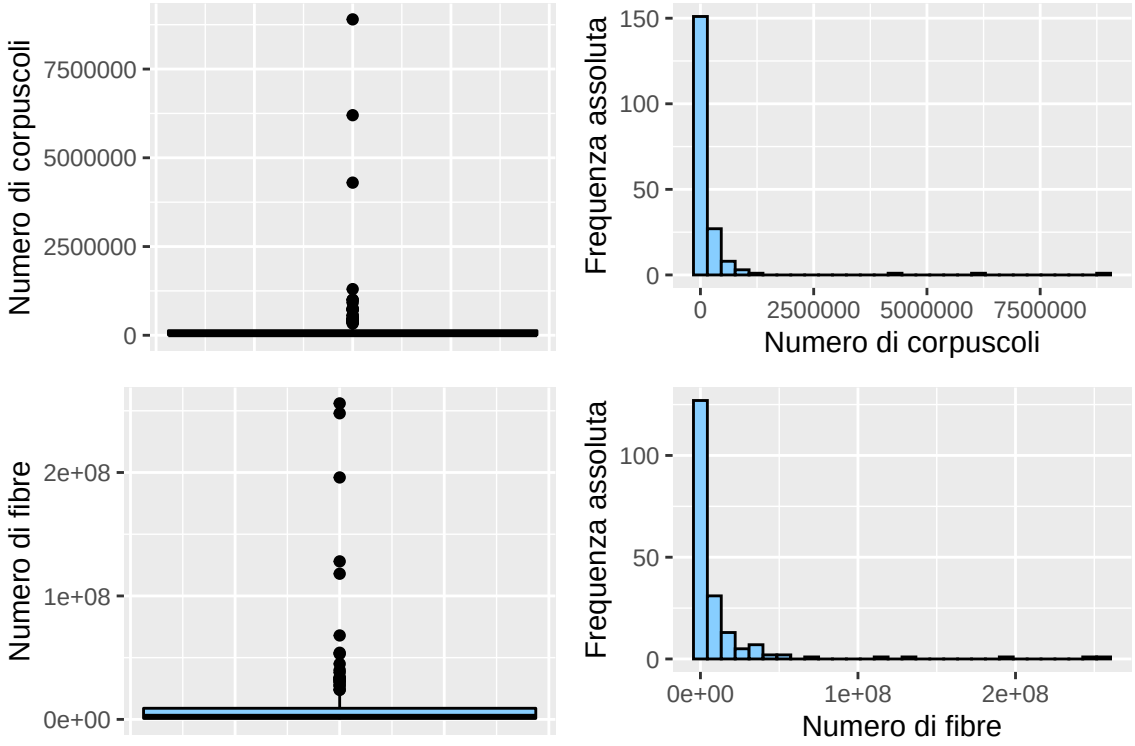

Figura 2.4.8: Distribuzioni del numero di fibre e di corpuscoli.

La quantità relativa al logaritmo numero di corpuscoli va da un minimo di 5.298 ad un massimo di 16.002; la media è 10.245 e la mediana è 10.204.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
log(corpuscoli)	5.298	8.854	10.2	10.25	11.78	16	2.17

Tabella 2.4.8: Statistiche di sintesi del numero di corpuscoli (scala logaritmica).

Come si può dedurre dai grafici e dalla vicinanza tra media e mediana, la trasformazione logaritmica della quantità sembra avere un andamento simile alla normale, risultato confermato dal test di Shapiro Wilk che con un p-value di 0.21 accetta l'ipotesi nulla di normalità.

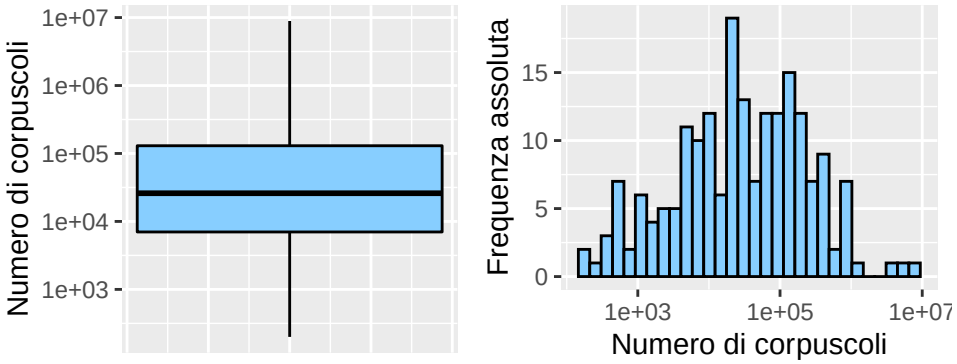

Figura 2.4.9: Distribuzioni del numero di corpuscoli (scala logaritmica).

La grandezza *fibre*, presa in scala logaritmica per motivi di praticità, va da un minimo di 11.29 ad un massimo di 19.36 log-fibre/grammo di tessuto secco; la media è 14.79 e la mediana è 14.6.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
log(fibre)	11.29	13.64	14.6	14.79	16.01	19.36	1.67

Tabella 2.4.9: Statistiche di sintesi del numero di fibre (scala logaritmica).

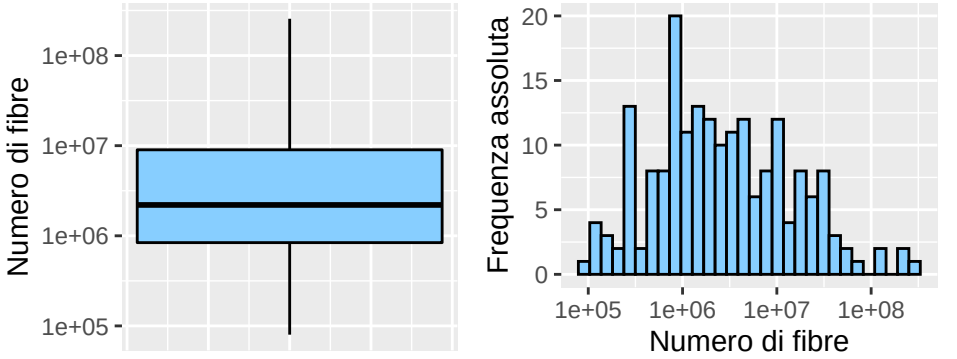

Figura 2.4.10: Distribuzioni del numero di fibre (scala logaritmica).

Come si può notare, pur avendo effettuato una trasformazione logaritmica, la distribuzione non appare strettamente normale; il test di Shapiro-Wilk conferma questa ipotesi, infatti si rifiuta l'ipotesi nulla di normalità al 5%, ma il p-value osservato è 0.03 quindi molto vicino alla soglia.
Analizzando gli intervalli di confidenza per il logaritmo del numero di corpuscoli e fibre, si nota che il logaritmo dell'ampiezza degli intervalli varia principalmente tra 8 e 12 per il primo, tra 14 a 16 per il secondo. È interessante notare la forte relazione lineare positiva tra stime ed ampiezza degli intervalli.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
log(Ampiezza.Corp)	4.605	8.854	9.903	10.082	11.408	15.253	1.78
log(Ampiezza.Fibre)	12.3	14.13	14.63	14.93	15.75	18.81	1.3

Tabella 2.4.10: Statistiche di sintesi dell'ampiezza degli intervalli di confidenza (scala logaritmica).

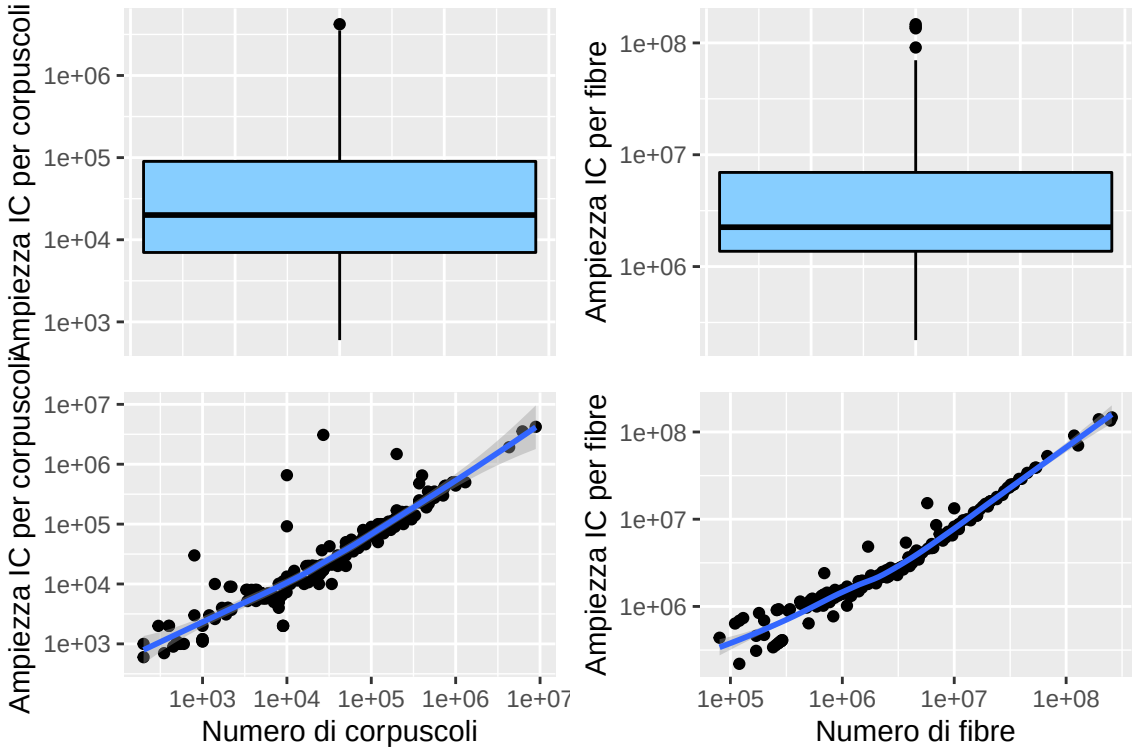

Figura 2.4.11: Distribuzioni dell'ampiezza degli intervalli su corpuscoli e fibre (scala logaritmica); relazioni con le stime.

È stato inoltre calcolato il numero di fibre di anfiboli ed il numero di fibre di crisotilo (denominandoli rispettivamente *fibre.anf* e *fibre.cris*), semplicemente moltiplicando il numero di fibre per la percentuale di anfiboli e di crisotilo. Come era facilmente intuibile, la loro distribuzione è nettamente asimmetrica a destra, come si può notare dai grafici sottostanti.

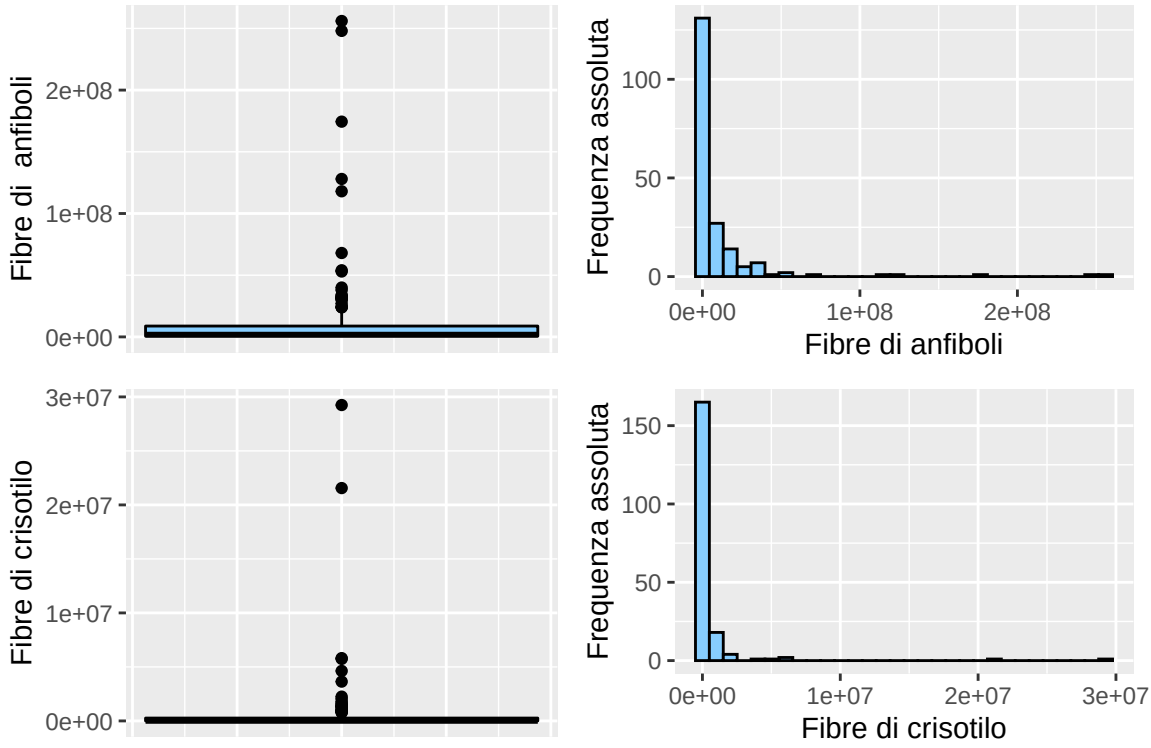

Figura 2.4.12: Distribuzioni del numero di fibre di anfiboli e crisotilo.

Di conseguenza è stata utilizzata una trasformazione logaritmica, ottenendo così *log.fibre.anf* e *log.fibre.cris*; dove il numero di un certo tipo di fibra era uguale a zero, anche il suo logaritmo è stato imputato pari a zero. La quantità relativa al logaritmo numero di fibre di anfiboli va da un minimo di 0 ad un massimo di 19.36 log-fibre/grammo di tessuto secco; la media è 14.43 e la mediana è 14.44.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
log(fibre.anf)	0	3.24	14.44	14.43	15.99	19.36	2.5

Tabella 2.4.11: Statistiche di sintesi delle fibre di anfiboli (scala logaritmica).

Come si può dedurre dalla vicinanza tra media e mediana, la trasformazione logaritmica della quantità oggetto di studio sembra avere un andamento normale, nonostante dai grafici si evinca una leggera asimmetria a destra. Il test di Shapiro Wilk con un p-value vicino a 0 rifiuta l'ipotesi nulla di normalità.

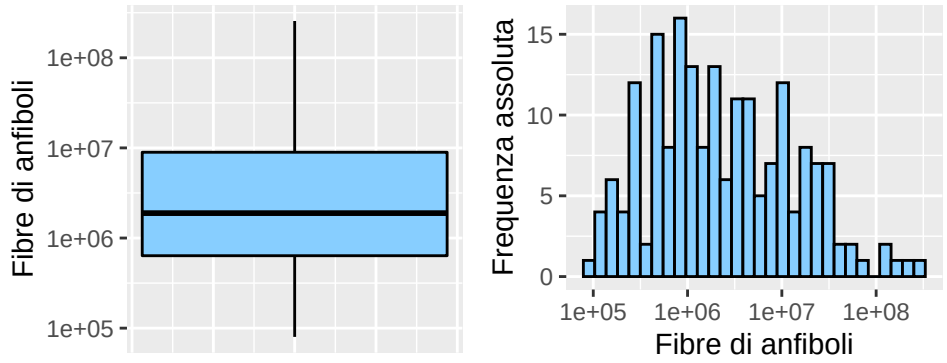

Figura 2.4.13: Distribuzioni del numero di fibre di anfiboli (scala logaritmica).

La quantità relativa al logaritmo numero di fibre di crisotilo va da un minimo di 0 ad un massimo di 17.191 log-fibre/grammo di tessuto secco; la media è 6 e la mediana è 0.

	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
log(fibre.cris)	0	0	0	6	12.484	17.191	6.38

Tabella 2.4.12: Statistiche di sintesi delle fibre di crisotilo (scala logaritmica).

Come si può dedurre dai grafici, la trasformazione logaritmica della quantità segue una distribuzione che sembra normale. Sono stati tolti i valori pari a zero in quanto oltre la metà dei soggetti (101 su 193) non presenta fibre di crisotilo, di conseguenza la distribuzione sarebbe stata pesantemente influenzata da questa alta frequenza di zeri. Togliendo i valori uguali a zero, il test di Shapiro Wilk rifiuta l'ipotesi nulla di normalità a livello 5%, ma il un p-value è pari a 0.04, quindi molto vicino alla soglia.

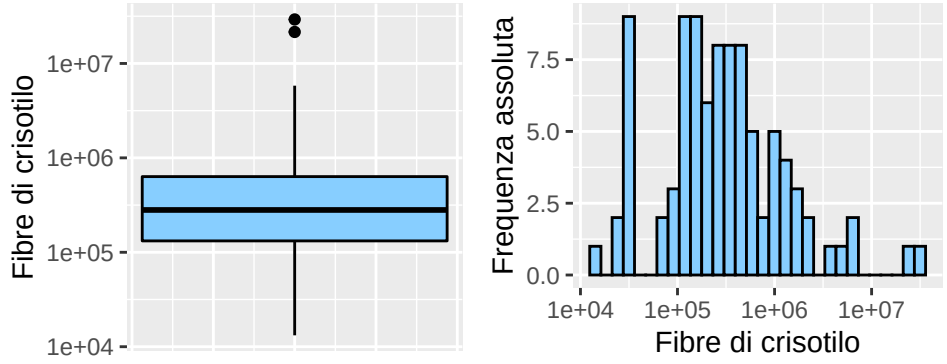

Figura 2.4.14: Distribuzioni delle fibre di crisotilo (scala logaritmica).

2.5 Analisi bivariate

In questo paragrafo vengono presentate le analisi descrittive bivariate tra la quantità oggetto di studio ovvero il numero di corpuscoli trovato e le altre caratteristiche presenti nel dataset;

sono stati utilizzati grafici e tabelle riassuntive.
Analizzando la relazione tra numero di fibre e numero di corpuscoli (entrambi i conteggi sono presi in scala logaritmica) si nota un'evidente relazione lineare positiva. Dal momento che, come precedentemente dimostrato, entrambe le distribuzioni sono normali, è stato opportuno sfruttare il coefficiente di correlazione di Pearson in quanto fornisce una descrizione completa dell'associazione. L'indice di correlazione risulta essere pari a 0.76; è risaputo infatti che un maggior numero di fibre di amianto nei polmoni comporta ad un maggior numero di corpuscoli che si formano.
Nella Figura 2.5.1 vengono distinti i casi in cui prevale un tipo di fibra rispetto ad un altro; i soggetti che risultano avere più fibre di anfiboli che di crisotilo sembrano avere anche un numero di corpuscoli maggiore. Entrambe le curve seguono un andamento lineare ed i coefficienti di correlazione sono rispettivamente 0.76 e 0.767.

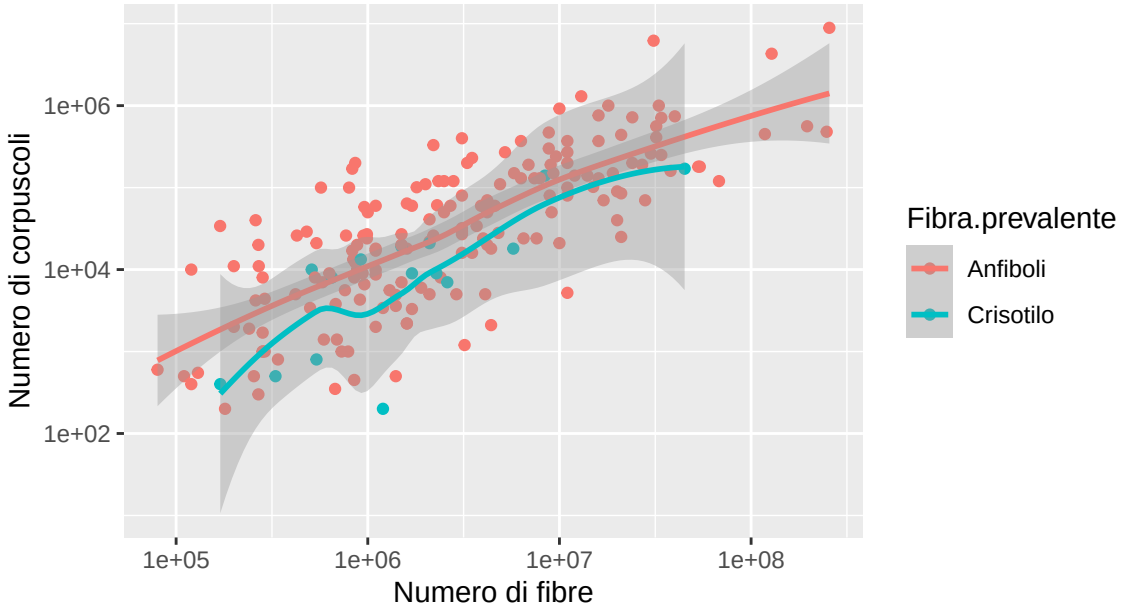

Figura 2.5.1: Relazione tra numero di corpuscoli e numero di fibre (scala logaritmica).

Successivamente è stata studiata la relazione tra numero di fibre a seconda del tipo di fibra e corpuscoli. Il grafico a sinistra mostra un'evidente relazione positiva ed un coefficiente di correlazione di Pearson pari a 0.64, mentre l'andamento del grafico a destra è vistosamente condizionato dalle 101 osservazioni dove non erano presenti fibre di crisotilo. La relazione lineare appare debole, con un coefficiente di correlazione di Pearson di 0.44, ottenuto togliendo i 101 valori senza fibre di crisotilo.

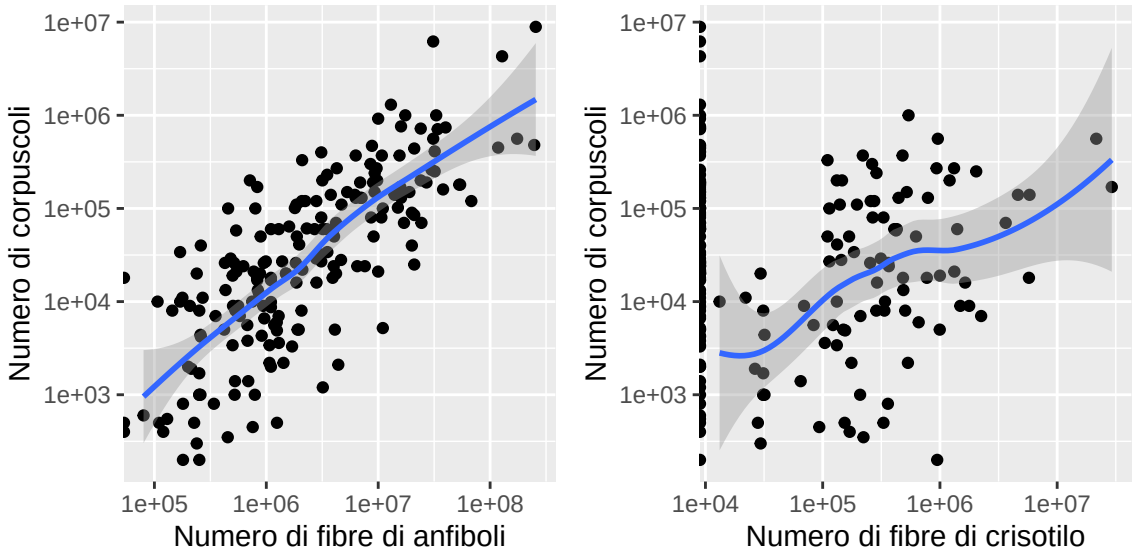

Figura 2.5.2: Relazione tra numero di corpuscoli e numero di fibre di anfiboli e crisotilo (scala logaritmica).

Nei boxplot che confrontano la distribuzione del logaritmo del numero di corpuscoli rispetto alle caratteristiche descrittive sulla provenienza e l'esposizione si nota che il numero maggiore di corpuscoli si trova nei soggetti residenti in Friuli, che sono stati in contatto con amianto per lavoro e più precisamente nei cantieri navali. Si nota inoltre che il numero di corpuscoli tende a crescere all'aumentare del livello d'esposizione e che non c'è un'evidente relazione con il tipo di fibra e con la malattia rilevata. Sembrano infine esserci più corpuscoli negli uomini rispetto alle donne, probabilmente perchè il loro lavoro era meno pesante ed erano sottoposte ad un minor carico di fibre.

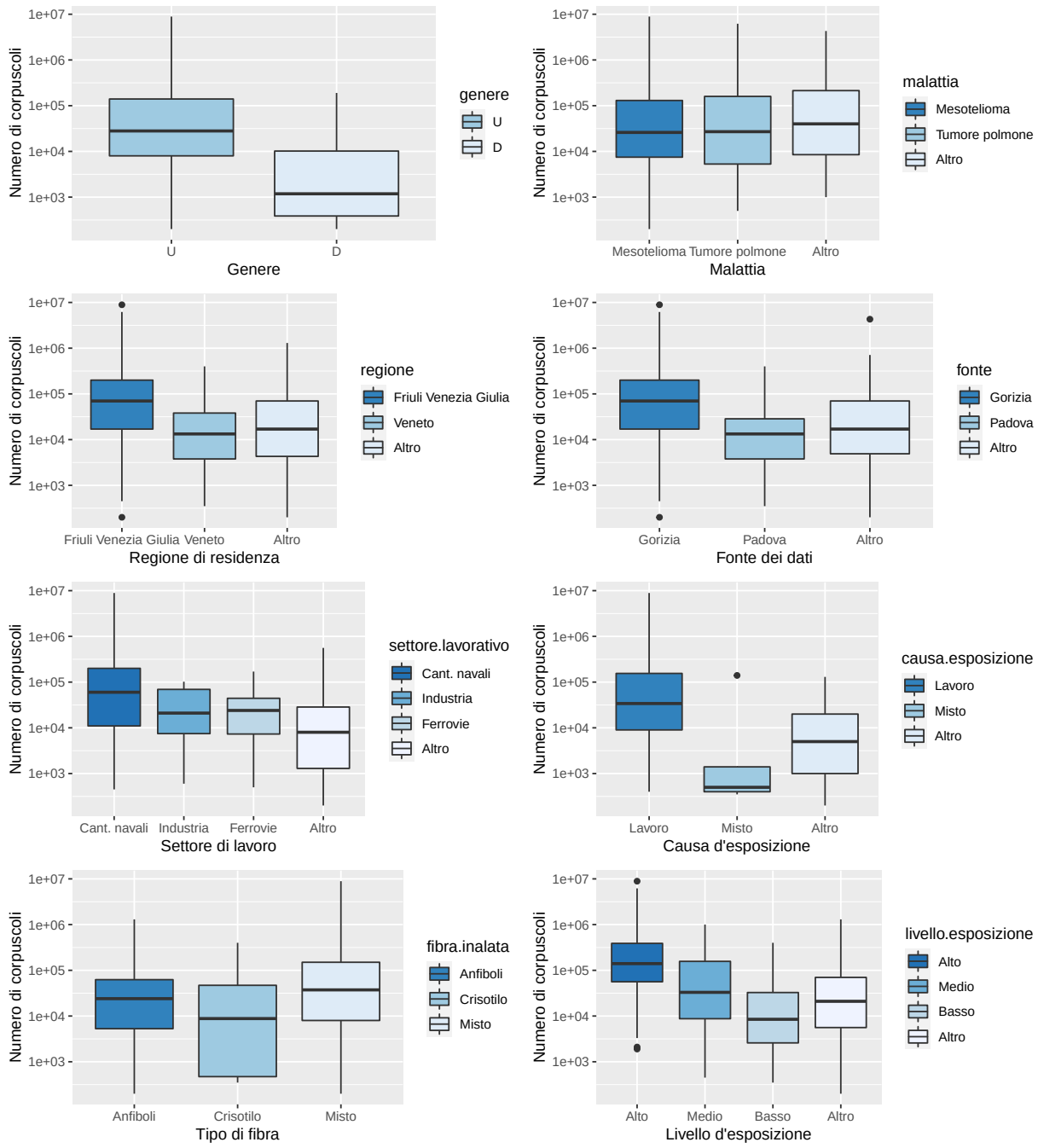

Figura 2.5.3: Relazioni tra numero di corpuscoli e caratteristiche descrittive sulla provenienza e sull'esposizione (scala logaritmica).

Secondo una prima analisi grafica il numero di corpuscoli non sembra avere una relazione lineare significativa con la durata di esposizione, con l'età di inizio esposizione e con il periodo di latenza. È importante ricordare che nonostante il 70% dei soggetti riporti un tipo di fibra inalato misto, quasi tutti riportano tracce di anfiboli ma solo la metà ha tracce di crisotilo e solo 15 soggetti su 193 hanno un contenuto di crisotilo maggiore rispetto agli

anfiboli. Inoltre, come già riportato in precedenza, le fibre di anfiboli, più frequenti nei soggetti contenuti dal dataset, hanno un tempo di smaltimento naturale (quantificabile in anni) molto più lungo rispetto alle fibre di crisotilo.
Dal primo grafico si nota una concentrazione maggiore di soggetti che hanno iniziato l'esposizione tra i 10 ed i 30 anni, in quanto solo 30 soggetti hanno cominciato ad essere esposti ad amianto dopo i 30 anni. Di conseguenza è naturale notare un andamento decrescente del logaritmo del numero di corpuscoli a partire dai 30 anni, in quanto i corpuscoli non hanno tempo a sufficienza per formarsi.
Il secondo grafico segue un andamento tutt'altro che lineare e l'interpretazione è complicata anche a causa dell'eteroschedasticità. Il logaritmo del numero di corpuscoli sembra diminuire quando la durata dell'esposizione è compresa tra 10 e 15 anni, per poi risalire ed abbassarsi quando la durata supera i 33 anni, forse per il fatto che esiste un numero "limite massimo" di corpuscoli di amianto fisicamente presenti all'interno dei polmoni.
Analizzando l'ultimo grafico, sembra avere un andamento crescente; nei primi decenni il logaritmo del numero di corpuscoli sembra abbastanza basso, forse per il fatto che lo smaltimento naturale è più efficace essendoci meno fibre. Successivamente cresce molto abbastanza velocemente verso i 32 anni di latenza e poi si stabilizza, probabilmente per lo stesso motivo del grafico precedente.
In tutti e tre i grafici è possibile notare l'ampia variabilità dei dati.

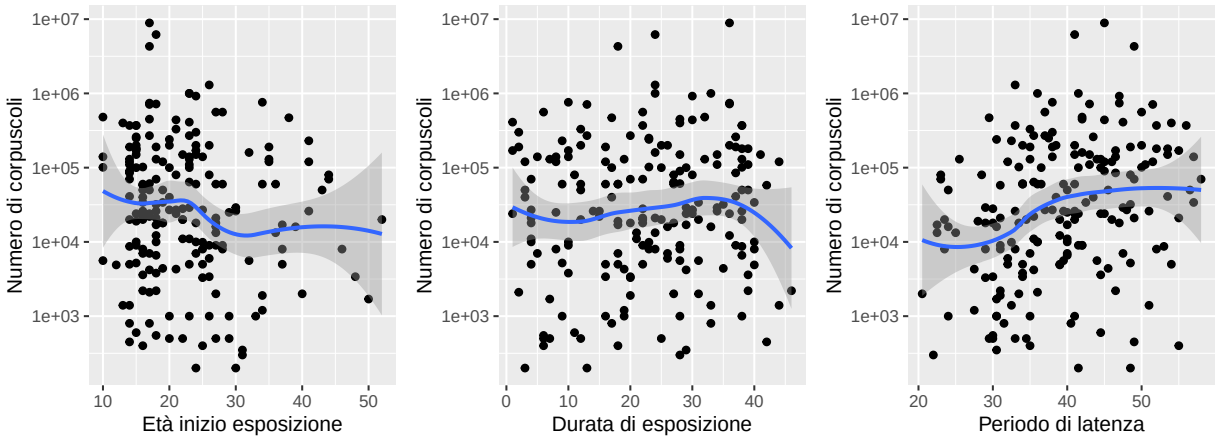

Figura 2.5.4: Relazioni tra numero di corpuscoli e le caratteristiche relative all'età di inizio dell'esposizione, alla durata dell'esposizione ed al periodo di latenza (scala logaritmica).

Come era prevedibile, i valori dei coefficienti di correlazione di Pearson sono piuttosto bassi, rispettivamente 0.073, -0.122 e 0.29. Le distribuzioni delle tre variabili non sono normali e le relazioni con il logaritmo del numero di corpuscoli non sono lineari, quindi il test non fornisce una descrizione completa dell'associazione.
Assieme ai boxplot vengono riportate delle tabelle riassuntive che descrivono la distribuzione del logaritmo del numero di corpuscoli a seconda delle esplicative discrete.

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Uomo	181	5.298	8.987	10.275	10.398	11.849	16.002	2.09
Donna	12	5.704	6.159	7.076	7.968	9.228	12.155	2.2

Tabella 2.5.1: Statistiche di sintesi del numero di corpuscoli a seconda del genere (scala logaritmica).

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Mesotelioma	127	5.298	8.92	10.166	10.177	11.775	16.002	2.139
Tumore polmone	55	6.215	8.574	10.204	10.315	11.981	15.64	2.215
Altro	11	7.09	8.987	10.597	10.706	12.272	15.274	2.53

Tabella 2.5.2: Statistiche di sintesi del numero di corpuscoli a seconda della malattia (scala logaritmica).

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Friuli	101	6.109	9.741	11.156	10.964	12.346	16.002	2.07
Veneto	55	5.858	8.237	9.496	9.328	10.82	14.947	1.96
Altro	37	5.298	8.366	9.741	9.656	11.156	14.078	2.11

Tabella 2.5.3: Statistiche di sintesi del numero di corpuscoli a seconda della regione di residenza (scala logaritmica).

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Gorizia	105	6.109	9.741	11.156	10.93	12.346	16.002	2.06
Padova	55	5.858	8.237	9.496	9.24	10.258	12.429	1.8
Altro	33	5.298	8.497	9.741	9.751	11.156	15.274	2.36

Tabella 2.5.4: Statistiche di sintesi del numero di corpuscoli a seconda della fonte (scala logaritmica).

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Cantieri navali	115	6.109	9.306	11.002	10.91	12.276	16.002	2.037
Ferrovie	16	6.215	8.898	10.086	9.622	10.698	12.044	1.845
Industria	11	6.397	8.919	9.952	10.139	11.29	14.947	2.3
Altro	51	5.298	7.167	8.987	8.972	10.258	13.236	1.96

Tabella 2.5.5: Statistiche di sintesi del numero di corpuscoli a seconda del settore lavorativo (scala logaritmica).

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Lavoro	167	5.991	9.105	10.597	10.54	11.951	16.002	2.05
Misto	5	5.858	5.991	6.215	7.432	7.244	11.849	2.53
Altro	21	5.298	7.09	8.517	8.589	9.903	11.775	1.92

Tabella 2.5.6: Statistiche di sintesi del numero di corpuscoli a seconda della causa d'esposizione (scala logaritmica).

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Anfiboli	51	5.991	8.574	10.086	10.018	11.209	14.947	2.12
Crisotilo	8	5.858	6.159	9.082	8.691	10.411	12.206	2.45
Misto	134	5.298	8.987	10.528	10.427	11.918	16.002	2.15

Tabella 2.5.7: Statistiche di sintesi del numero di corpuscoli a seconda del tipo di fibra inalato (scala logaritmica).

	N	Min.	1ºQ.	Mediana	Media	3ºQ.	Max.	S.d.
Alto	32	7.55	10.94	11.85	11.71	12.87	16.002	2.061
Medio	80	6.109	8.987	10.609	10.425	12.101	14.947	2.095
Basso	34	5.858	7.858	9.046	9.053	10.394	12.206	1.852
Altro	47	5.298	8.631	9.952	9.811	11.156	14.078	2.025

Tabella 2.5.8: Statistiche di sintesi del numero di corpuscoli a seconda del livello d'esposizione (scala logaritmica).

Oltre all'analisi esplorativa tramite i boxplot e le tabelle riassuntive, per verificare se esiste o meno una differenza significativa di numero di corpuscoli a seconda delle varie esplicative sono stati utilizzati diversi test a seconda della natura delle variabili.
Per le variabili *genere, malattia, regione, fonte, settore.lavorativo* e *livello.esposizione* è stato scelto un approccio parametrico, essendo verificate le condizioni necessarie ovvero gruppi sufficientemente numerosi, normalità del logaritmo del numero di corpuscoli per ogni gruppo ed omoschedasticità. La normalità è stata verificata tramite il test di Shapiro-Wilk, l'omoschedasticità grazie al test di Levene, l'ipotesi di uguaglianza in media della risposta nei vari gruppi con l'ANOVA ad una via ed i confronti multipli con il metodo di Holm, per variabili con più di due modalità, vale a dire *malattia*, *regione*, *fonte*, *settore.lavorativo* e *livello.esposizione*.

Come si può facilmente notare dai test sottostanti, possono essere accettate le assunzioni di normalità ed omoschedasticità delle distribuzioni del numero di corpuscoli tra uomini e donne. Dall'ANOVA, con un p-value minore di 0.001, si conclude che le due distribuzioni sono statisticamente differenti in media.

	Test statistic	P value
Uomini	0.9913	0.3452
Donne	0.8982	0.1505

Tabella 2.5.9: Test di Shapiro-Wilk del numero di corpuscoli per il genere (scala logaritmica).

Test statistic	P value
0.2871	0.5927

Tabella 2.5.10: Test di Levene del numero di corpuscoli per il genere (scala logaritmica).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genere	1	70.03	70.03	16.19	8.247e-05
Residuals	191	826.1	4.325	NA	NA

Tabella 2.5.11: ANOVA del numero di corpuscoli per il genere (scala logaritmica).

Analizzando la malattia dei soggetti, una volta verificate le dovute assunzioni (si veda l'Appendice A per i risultati completi delle successive analisi), si conclude con un p-value di 0.6758 che i tre gruppi sono statisticamente uguali in media, risultato deducibile anche dai boxplot.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
malattia	2	3.689	1.844	0.3927	0.6758
Residuals	190	892.5	4.697	NA	NA

Tabella 2.5.12: ANOVA del numero di corpuscoli per la malattia (scala logaritmica).

Il test di Holm conferma il risultato precedente, con tutti i p-value pari ad 1.

	Altro	Mesotelioma
Mesotelioma	1	NA
Tumore polmone	1	1

Tabella 2.5.13: Confronti multipli del numero di corpuscoli per la malattia (scala logaritmica).

Per la regione di residenza, verificate le assunzioni, si nota che con un p-value vicino a 0 che viene rifiutata l'ipotesi di uguaglianza delle tre medie.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
regione	2	110.7	55.34	13.39	3.638e-06
Residuals	190	785.5	4.134	NA	NA

Tabella 2.5.14: ANOVA del numero di corpuscoli per la regione di residenza (scala logaritmica).

Dall'output del metodo di Holm è evidente che i due gruppi statisticamente differenti sono Friuli Venezia Giulia-Altro (p-value=0.0027) e Friuli Venezia Giulia-Veneto (p-value<0.001).

	Altro	Friuli Venezia Giulia
Friuli Venezia Giulia	0.002746	NA
Veneto	0.3843	8.357e-06

Tabella 2.5.15: Confronti multipli del numero di corpuscoli per la regione di residenza (scala logaritmica).

Riguardo la fonte dei dati, verificate le assunzioni, si conclude che i tre gruppi sono statisticamente differenti in media, con un p-value vicino a 0.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fonte	2	108.9	54.43	13.14	4.535e-06
Residuals	190	787.3	4.144	NA	NA

Tabella 2.5.16: ANOVA del numero di corpuscoli per la fonte (scala logaritmica).

Il test post-hoc restituisce dei risultati simili al caso precedente. Infatti Gorizia rappresenta il Friuli Venezia Giulia e Padova il Veneto; anche guardando i boxplot si può osservare la similitudine delle distribuzioni.

	Altro	Gorizia
Gorizia	0.003574	NA
Padova	0.4645	1.107e-05

Tabella 2.5.17: Confronti multipli del numero di corpuscoli per la fonte (scala logaritmica).

Studiando il settore lavorativo dei soggetti, verificate le consuete assunzioni, si decide con un p-value vicino 0 di rifiutare l'ipotesi nulla di uguaglianza in media.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
settore.lavorativo	3	138	46.01	11.47	6.068e-07
Residuals	189	758.1	4.011	NA	NA

Tabella 2.5.18: ANOVA del numero di corpuscoli per il settore lavorativo (scala logaritmica).

Come prevedibile dai boxplot, gli unici due gruppi statisticamente differenti in media (ad un livello di confidenza del 95%) sono "Cantieri navali" ed "Altro".

	Altro	Cant. navali	Industria
Cant. navali	3.29e-07	NA	NA
Industria	0.7736	0.2625	NA
Ferrovie	0.7736	0.09868	0.9127

Tabella 2.5.19: Confronti multipli del numero di corpuscoli per il settore lavorativo (scala logaritmica).

Infine è stato analizzato il livello d'esposizione: verificate le opportune assunzioni, si nota che si rifiuta l'ipotesi nulla di uguaglianza in media dei quattro gruppi con un p-value vicino

a 0.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
livello.esposizione	3	123.2	41.08	10.05	3.571e-06
Residuals	189	772.9	4.089	NA	NA

Tabella 2.5.20: ANOVA del numero di corpuscoli per il livello d'esposizione (scala logaritmica).

Con il confronto a coppie, si vede che si accetta l'ipotesi nulla di uguaglianza in media solo tra i gruppi "Altro"-"Basso" e "Altro"-"Medio", con un p-value di 0.2777.

	Alto	Altro	Basso
Altro	0.0002074	NA	NA
Basso	2.856e-06	0.2777	NA
Medio	0.004727	0.2777	0.0121

Tabella 2.5.21: Confronti multipli del numero di corpuscoli per il livello d'esposizione (scala logaritmica).

Le variabili *causa.esposizione* e *fibra.inalata* sono state trattate in maniera differente in quanto le rispettive modalità "Misto" e "Crisotilo" contengono solo 5 ed 8 osservazioni e dunque la normalità è difficile da verificare. Di conseguenza per valutare se c'è una differenza statisticamente significativa del logaritmo del numero di corpuscoli nei vari gruppi è stato utilizzato il test ANOVA per ranghi a una via di Kruskal-Wallis ed il test di Mann-Whitney per ogni coppia per i confronti multipli.
Per quanto riguarda *causa.esposizione*, il primo test rifiuta con forza l'ipotesi nulla di uguaglianza in mediana, il secondo riporta che si può accettare H_0 solo per la coppia "Misto"-"Altro", con un p-value pari a 0.3286.

Test statistic	df	P value
20.16	2	4.188e-05 * * *

Tabella 2.5.22: ANOVA per ranghi a una via di Kruskal-Wallis del numero di corpuscoli per la causa d'esposizione (scala logaritmica)[1].

[1] $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

	Altro	Lavoro
Lavoro	0.0004086	NA
Misto	0.3286	0.02113

Tabella 2.5.23: Confronti multipli del numero di corpuscoli per la causa d'esposizione (scala logaritmica).

Studiando infine il tipo di fibra inalato, si nota che il primo test accetta l'ipotesi nulla di uguaglianza in mediana con un p-value uguale a 0.1683, risultato ribadito dal secondo test che accetta H_0 per ogni coppia.

Test statistic	df	P value
3.564	2	0.1683

Tabella 2.5.24: ANOVA per ranghi a una via di Kruskal-Wallis del numero di corpuscoli per il tipo di fibra (scala logaritmica).

	Anfiboli	Crisotilo
Crisotilo	0.4744	NA
Misto	0.4744	0.4744

Tabella 2.5.25: Confronti multipli del numero di corpuscoli per il tipo di fibra (scala logaritmica).

3 Modelli di regressione

Questo capitolo è dedicato alla costruzione del modello di regressione con il quale è stata quantificata l'influenza delle variabili esplicative sul fenomeno oggetto di studio, ovvero la formazione di corpuscoli di amianto in pazienti che hanno subito in passato un'esposizione. Vengono descritte le ragioni ed il procedimento che hanno portato al modello conclusivo e la sua interpretazione e bontà di adattamento. Si è voluto inoltre descrivere accuratamente le motivazioni delle scelte prese, al fine di risolvere alcuni problemi metodologici.

3.1 Teoria dei minimi quadrati ordinari

Il metodo dei minimi quadrati ordinari (OLS, ovvero Ordinary Least Squares) [17] è una tecnica per stimare i parametri ignoti di un modello di regressione lineare; permette di trovare una funzione che sia più vicina possibile ad un insieme di dati. Tale funzione deve essere quella che minimizza la somma dei quadrati delle distanze tra i dati osservati e quelli della curva che rappresenta la medesima funzione. Si vede un esempio nel caso della regressione lineare semplice, in cui ci sono solo due variabili coinvolte: la variabile risposta Y ed un'unica variabile esplicativa X. Il modello è scritto nella forma

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

dove β_1 e β_2 sono i due parametri di regressione ignoti, x_i sono le realizzazioni di X ed ε_i sono i termini d'errore. Il modello poggia sulle assunzioni di linearità, ovvero $E(Y_i|X_i = x_i) = \beta_1 + \beta_2 x_i$, di omoschedasticità, cioè $V(Y_i|X_i = x_i) = \sigma^2 > 0$ e media nulla, normalità, indipendenza ed omoschedasticità degli errori, vale a dire $\varepsilon_i \sim N(0, \sigma^2)$ indipendenti, $i = 1, ..., n$. Si può ottenere una stima dei due parametri di regressione tramite il metodo dei minimi quadrati ordinari, ovvero trovando i due parametri che minimizzano la quantità

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - (\beta_1 + \beta_2 x_i))^2$$

che è una misura della distanza tra i valori osservati y_i ed una generica retta $\beta_1 + \beta_2 x_i$, in quanto è la sommatoria degli scarti al quadrato. Minimizzando tale quantità si ottiene la stima dei minimi quadrati ordinari di (β_1, β_2), ovvero

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{(\beta_1,\beta_2)\in R^2}{argmin} \sum_{i=1}^{n} (y_i - (\beta_1 + \beta_2 x_i))^2 = (\overline{y} - \hat{\beta}_2 \overline{x}, \ \frac{\sigma_{xy}}{\sigma_x^2})$$

dove $\overline{y}$ e $\overline{x}$ sono le due medie aritmetiche, σ_{xy} è la covarianza di x ed y e σ_x^2 la varianza di x. Applicando tale metodo alla regressione lineare multipla, in cui ci sono $p > 1$ variabili esplicative, occorre minimizzare la funzione del vettore di parametri incogniti β di lunghezza p

$$S(\beta) = (y - X\beta)^T (y - X\beta)$$

dove y è il vettore di valori osservati di lunghezza n ed X la matrice di regressione di dimensione $n \times p$. La stima dei minimi quadrati del vettore β risulta essere

$$\hat{\beta} = \underset{\beta}{argmin}\ S(\beta) = (X^T X)^{-1} X^T y$$

3.2 Applicazione del modello OLS

Il primo approccio utilizzato per valutare l'influenza delle variabili esplicative sul fenomeno d'interesse è quello di un modello di regressione ai minimi quadrati ordinari.
Si è scelto di non candidare come variabili esplicative del modello tutte quelle trattate nel capitolo precedente, per un motivo ben preciso. Lo scopo principale di questa tesi, più di andare a quantificare la mera correlazione tra variabili esplicative e fenomeno d'interesse, è quello di andare a valutare quali sono i fattori che influenzano il processo di formazione dei corpuscoli in pazienti esposti ad amianto. Di conseguenza, mentre nel Capitolo 2 è stata trattata la relazione marginale tra esplicative e risposta, in questo capitolo dedicato ai modelli si è voluto quantificare l'influenza congiunta di alcuni fattori sulla formazione di corpuscoli di amianto.
Nella selezione delle variabili da inserire nel modello sono state fatte alcune scelte logiche al fine di evitare multicollinearità ed informazioni ripetute.
Un primo caso è quello della quantità ed il tipo di fibre; per evitare ripetizioni di informazioni tra *log.fibre*, *anfiboli*, *crisotilo* e *fibra.inalata* è stato scelto di non includere quest'ultima variabile.
Nonostante non siano state considerate nel modello poichè sono più di natura descrittiva dei soggetti presenti nel dataset, si è visto che alcune variabili contengono grossomodo le stesse informazioni.
Si è visto, ad esempio che le variabili *fonte* e *regione* sono molto correlate tra loro. Per valutarne la dipendenza è stato applicato il test χ^2 di Pearson, che permette di valutare l'associazione tra le due variabili discrete. Denominate A e B le due variabili, il test verifica il test d'ipotesi

$$\begin{cases} H_0 : A\ e\ B\ sono\ indipendenti \\ H_1 : A\ e\ B\ non\ sono\ indipendenti \end{cases}$$

Chiamate $A_1, ...A_i, ...A_I$ e $B_1, ..., Bj, ..., B_J$ le relative modalità, si costruisce la tabella di frequenza per le n osservazioni e viene chiamata f_{ij} la generica frequenza assoluta osservata per le modalità i di A e j di B. Le frequenze attese in caso di indipendenza sono $f^*_{ij} = \frac{f_{i+} f_{+j}}{n}$, dove f_{i+} ed f_{+j} sono le frequenze marginali di riga e di colonna. La statistica test è calcolata come

$$X^2 = \sum_i \sum_j \frac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}}$$

Il test perde potenza se una o più frequenze assolute è minore di 5. Per n sufficientemente grande, si la distribuzione nulla di X^2 è una $\chi^2_{(I-1)(J-1)}$. Applicando tale test a *fonte* e

regione, con un p-value vicino a zero si decide di rifiutare l'ipotesi nulla di indipendenza.
Analizzando la distribuzione congiunta di *fonte* e *settore.lavorativo*, si è notato che la maggior parte dei soggetti provenienti da Gorizia e Padova ha lavorato rispettivamente nei cantieri navali e nelle ferrovie. Applicando il test di indipendenza alle due variabili si è ottenuto il medesimo risultato di prima.
Non è stato possibile applicare questo test a *causa.esposizione* perchè la modalità "Misto" contiene solo 5 osservazioni.
Non sono state prese in considerazione neanche le variabili *genere*, *malattia* ed *intervento*, in quanto sono di poco interesse per analizzare e capire il processo evolutivo delle fibre di amianto in corpuscoli.
Nell'elenco sottostante vengono raggruppate tutte le esplicative candidate ad entrare nel modello:

- variabili relative alla quantità ed al tipo di fibre inalate: *log.fibre*, *anfiboli*, *crisotilo*, rispettivamente il logaritmo del numero di fibre e le percentuali di anfiboli e crisotilo presenti. Per motivi logici, queste ultime due variabili sono state considerate come dummy, che valgono 1 se la percentuale è inferiore a 90% e 0 altrimenti;
- variabili temporali: l'età di inizio esposizione chiamata *età.inizio*, divisa in tre classi d'età (<20 anni, 20-30 anni, >30 anni), la durata dell'esposizione chiamata *anni.esp*, divisi in quattro classi d'età (<10 anni, 10-20 anni, 20-30 anni, >30 anni) ed il periodo di latenza chiamato *latenza*, considerata come dummy, che vale 1 se la latenza è inferiore ai 33 anni e 0 altrimenti;
- variabile relative al periodo di esposizione, vale a dire *livello.esposizione*, divisa in quattro classi (Alto, Medio, Basso, Altro).

Come già ribadito più volte, al numero di corpuscoli Y è stata applicata una trasformazione logaritmica per assumerne la normalità. Dal momento che Y non è altro che la media geometrica di variabili log-normali, si può affermare che

$$Z = log(Y) \sim N(\mu_Z, \sigma_Z^2)$$

quindi Z è la variabile risposta presa in considerazione nei modelli successivi, che rappresenta il logaritmo del numero di corpuscoli.
È stato utilizzato un approccio forward, che consiste nell'includere nel modello le variabili esplicative che presentano il contributo maggiormente significativo, ossia il livello di significatività osservato minore per il test di nullità del coefficiente di regressione corrispondente (e minore del livello α fissato).
In particolare è stata usata la funzione *add*1 di R, che permette di confrontare tutti i possibili modelli che possono essere costruiti aggiungendo una variabile esplicativa; è stato usato il test F, adatto per modelli lineari.
Per valutare la selezione del modello migliore si è guardato a più fattori: oltre alla significatività delle variabili esplicative, l'AIC ed il BIC da minimizzare ed il coefficiente di

determinazione corretto $\overline{R^2}$ da massimizzare.

Il modello finale scelto è il seguente ed include le variabili *log.fibre*, *anfiboli* e *latenza*:

$$\log(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

dove $y_i = corpuscoli_i$, $x_{i2} = log.fibre_i$, $x_{i3} = \begin{cases} 1 & se \quad anfiboli < 90\% \\ 0 & altrimenti \end{cases}$,

$x_{i4} = \begin{cases} 1 & se \quad latenza < 33 \quad anni \\ 0 & altrimenti \end{cases}$ ed $\varepsilon_i \sim N(0, \sigma^2)$.

	Dependent variable:
	y
log.fibre	0.909^{***}
	(0.062)
I(anfiboli <90)	-0.491^{**}
	(0.213)
I(latenza <33)	-0.697^{***}
	(0.232)
Constant	-2.898^{***}
	(0.951)
Observations	193
R^2	0.620
Adjusted R^2	0.614
Akaike Inf. Crit.	667.381
Bayesian Inf. Crit.	683.694
Residual Std. Error	1.343 (df = 189)
F Statistic	102.722^{***} (df = 3; 189)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabella 3.2.1: Modello finale OLS.

La Figura 3.2.1 [20] serve a chiarire graficamente l'effetto delle esplicative sulla risposta.
La variabile più significativa è senza dubbio *log.fibre*; a parità di altre variabili, un aumento dell'1% del numero di fibre corrisponde ad un aumento dello 0.91% di corpuscoli (IC 95%: (0.78, 1.03)). Sempre a parità di altre variabili, la presenza di una percentuale di fibre inferiore al 90% riduce il numero di corpuscoli del 49.1% (IC 95%: (48.66, 49.5)) mentre latenze inferiori a 33 anni riducono la percentuale di corpuscoli del 69.7% (IC 95%: (69.26, 70.16)).
Gli intervalli di confidenza al 95% sulle stime dei coefficienti sono stati calcolati tramite la formula $(\hat{\beta} \pm z_{1-\frac{\alpha}{2}} \cdot se(\hat{\beta}))$, dove *se* è lo standard error della stima.

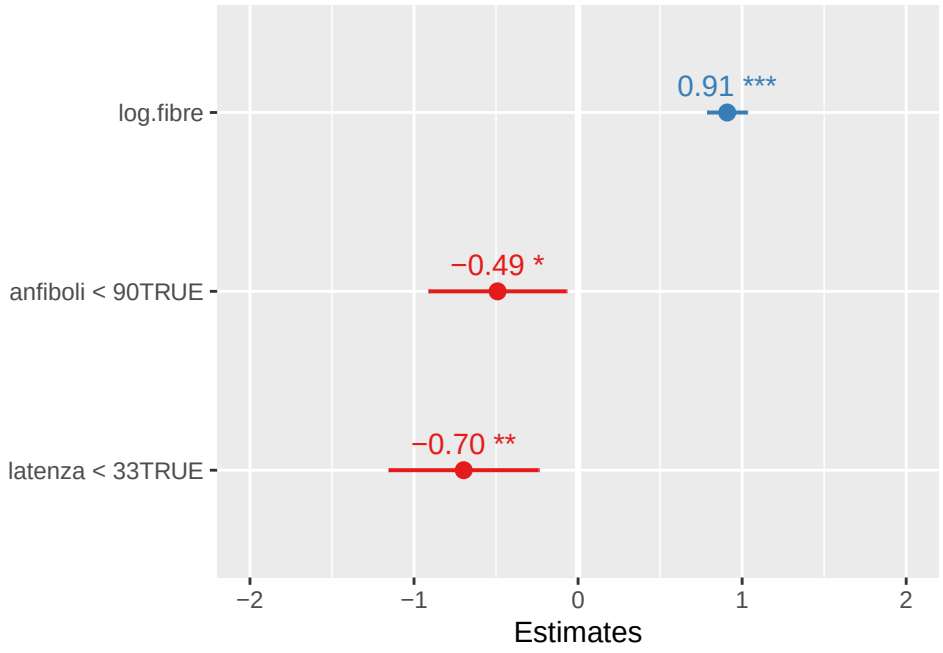

Figura 3.2.1: Rappresentazione grafica delle stime del modello OLS.

Come si può notare dalla Tabella 3.2.1, tutte le esplicative sono significative ad un livello del 95%; il coefficiente di determinazione R^2 sembra soddisfacente, pari a 0.6198, così come il coefficiente di determinazione corretto $\overline{R^2}$, uguale a 0.6138.
Vengono inoltre riportati nella Figura 3.2.2 quattro grafici relativi alla bontà di adattamento del modello.
Il grafico in alto a sinistra mostra i valori previsti in ascissa contro i residui del modello in ordinata, quello in basso a sinistra i valori previsti in ascissa contro la radice quadrata dei residui standardizzati del modello in ordinata. In entrambi i casi il grafico di dispersione per condurre ad un buon adattamento del modello deve seguire un andamento casuale e non sistematico. Il grafico in alto a destra è un grafico quantile-quantile che serve a verificare l'assunzione di normalità dei residui standardizzati (si veda il paragrafo 4.3); se sono normali si distribuiscono lungo una retta. In ultimo, il grafico in basso a destra è relativo alle distanze di Cook, che misurano l'effetto causato sul modello dalla rimozione di un determinato valore; nell'analisi con il metodo dei minimi quadrati ordinari può essere usata per indicare punti ad alta influenza.
I grafici della bontà di adattamento del modello portano a buone considerazioni sulla normalità, meno sull'omoschedasticità. Il modello infatti appare difettoso da questo punto di vista, forse a causa di alcuni valori anomali.

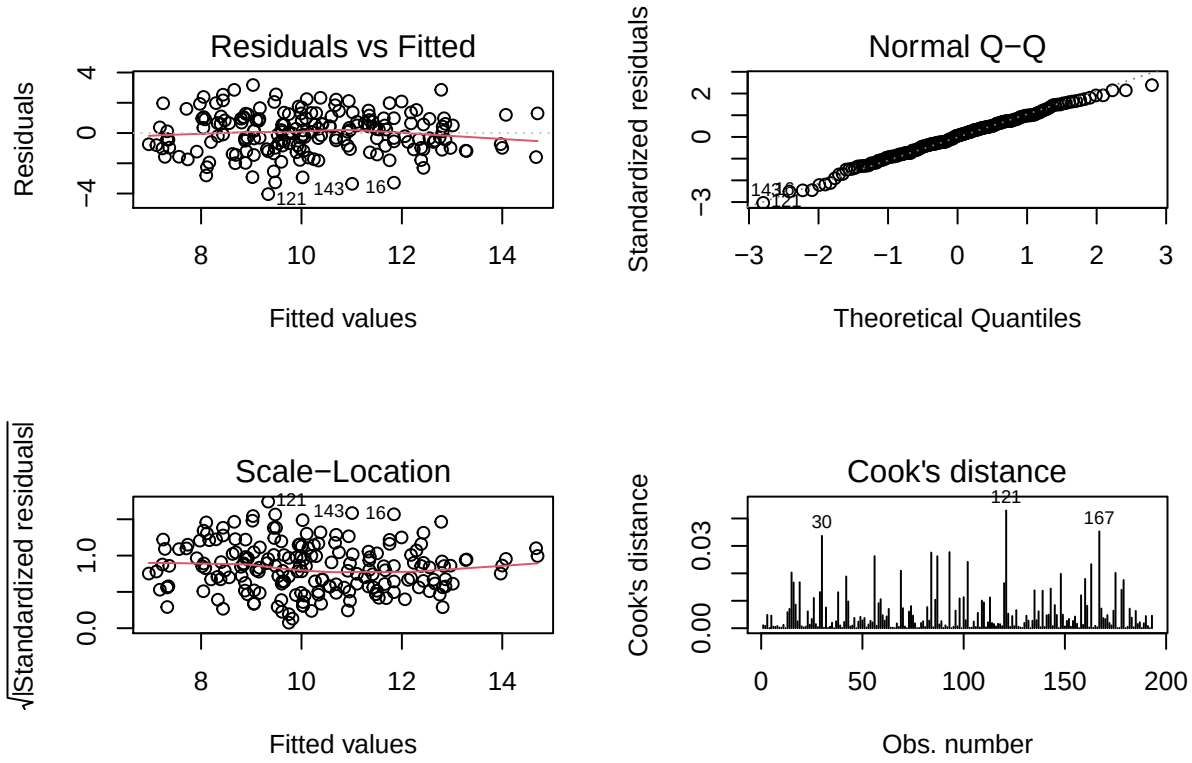

Figura 3.2.2: Adattamento del modello OLS.

Nel corso della costruzione del modello inserendo una alla volta le variabili, è stata tenuta in considerazione l'eventuale multicollinearità. Si tratta di un fenomeno in cui una o più coppie di variabili esplicative sono molto correlate tra di loro, rendendo la matrice dei regressori X non più a rango pieno e quindi non invertibile. Ciò può portare a degli standard error eccessivamente elevati, a delle stime peggiori ed un peggiore adattamento dei residui. Le variabili collineari non forniscono informazioni aggiuntive e risulta complicato individuare l'effetto che ciascuna di esse ha sulla variabile risposta. Un semplice controllo empirico può tenere sotto controllo questo eventuale problema, osservando se con l'aggiunta di una nuova esplicative si hanno standard error alti o outliers nei residui del modello.

3.3 Teoria dei minimi quadrati pesati

Dopo aver effettuato l'analisi tramite un modello OLS, è stato utilizzato anche un secondo tipo di modello, vale a dire un modello di regressione pesata.
Il motivo di questa scelta è che oltre ad avere una stima del numero di corpuscoli per ogni soggetto, vale a dire la quantità *corpuscoli*, è disponibile anche una stima dell'incertezza di tale valore, ovvero l'intervallo di ampiezza *Ampiezza.Corp=corpuscoli.UL-corpuscoli.LL*, che può rappresentare un intervallo di confidenza al 95% per *corpuscoli*. Di conseguenza è opportuno includere nel modello questa informazione aggiuntiva riguardante la variabilità

delle stime.
Un secondo motivo è il fatto che in alcuni casi può non essere verificata l'assunzione di omoschedasticità, ovvero la varianza degli errori non è costante. Nella maggior parte dei casi si nota quanto il grafico di dispersione della risposta contro i valori previsti o contro un'esplicativa non segue un andamento casuale, ma sistematico.
Può essere quindi utile assegnare a qualche osservazione un peso diverso rispetto ad altre; si interviene di conseguenza sulla varianza degli errori. In una situazione di omoschedasticità, nel caso di regressione multivariata, il vettore degli errori ε di lunghezza n segue una distribuzione normale multivariata a media nulla e matrice di varianze e covarianze diagonale con σ^2 sulla diagonale principale, ovvero $\varepsilon \sim N_n(\underline{0}, \sigma^2 I)$, con

$$E(\varepsilon) = \underline{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad e \qquad V(\varepsilon) = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

In caso di eteroschedasticità, la varianza dell'errore non è costante ma varia da unità ad unità a seconda di una certa funzione dei valori osservati $h(x_i)$, ovvero $V(\varepsilon_i) = \sigma^2 h(x_i)$; la matrice di varianze e covarianze del termine d'errore diventa quindi

$$V(\varepsilon_i) = \begin{bmatrix} \sigma^2 h(x_1) & 0 & \cdots & 0 \\ 0 & \sigma^2 h(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 h(x_n) \end{bmatrix}$$

È necessario quindi intervenire moltiplicando $V(\varepsilon_i)$ per una matrice, che sarà una matrice diagonale con il reciproco di $h(x_i)$ sulla diagonale principale, in modo tale da avere

$$V(\varepsilon_i^*) = \begin{bmatrix} \sigma^2 h(x_1) & 0 & \cdots & 0 \\ 0 & \sigma^2 h(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 h(x_n) \end{bmatrix} \cdot \begin{bmatrix} 1/h(x_1) & 0 & \cdots & 0 \\ 0 & 1/h(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/h(x_n) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

La matrice moltiplicata è detta W ed è la matrice diagonale di dimensione $n \times n$ che contiene sulla diagonale principale i pesi $\frac{1}{h(x_i)}$ assegnati a ciascuna osservazione e quindi la sua inversa sarà

$$W^{-1} = \begin{bmatrix} h(x_1) & 0 & \cdots & 0 \\ 0 & h(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h(x_n) \end{bmatrix}$$

Riassumendo, si passa da ε_i con varianza $V(\varepsilon_i) = \sigma^2 W^{-1}$ ad ε_i^* con varianza $V(\varepsilon_i^*) = \sigma^2 W^{-1} W = \sigma^2$.

Come si è visto è stata applicata una trasformazione a ε_i, così come a tutto il modello, che è diventato

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + ... + \beta_p x_{ip}^* + \varepsilon_i^* \qquad i = 1, ...n$$

con $y_i^* = \frac{y_i}{\sqrt{h(x_i)}}$, $\quad x_{ij}^* = \frac{x_{ij}}{\sqrt{h(x_i)}}$, $\quad \varepsilon_i^* = \frac{\varepsilon}{\sqrt{h(x_i)}}$, in modo tale da avere

$$V(y_i^*) = V(\beta_1 x_{i1}^* + ... + \beta_p x_{ip}^* + \varepsilon_i^*) = V(\varepsilon_i^*) = V(\frac{\varepsilon_i}{\sqrt{h(x_i)}}) = (\frac{1}{\sqrt{h(x_i)}})^2 \cdot V(\varepsilon_i) = \frac{1}{h(x_i)}\sigma^2 h(x_i) = \sigma^2$$

Esistono diversi modi per calcolare i pesi, in questa tesi ne sono stati considerati alcuni.
Questo metodo è chiamato metodo dei minimi quadrati pesati (WLS, ovvero Weighted Least Squares [21], in italiano "minimi quadrati ponderati"). Anzichè minimizzare la somma delle distanza al quadrato, è quindi più opportuno minimizzare la somma delle distanza al quadrato pesate, ossia

$$S_w(\beta_1, ..., \beta_p) = \sum_{i=1}^{n} w_i (y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}))^2$$

ottenendo la stima dei minimi quadrati pesati, ovvero

$$\hat{\beta}_{WLS} = \underset{\beta \in R^P}{argmin} \; S_w(\beta) = (X^T W X)^{-1} X^T W y$$

dove $\beta = (\beta_1, ..., \beta_p)$ è il vettore dei parametri, $\hat{\beta}$ è il vettore dei parametri stimati di lunghezza p, X è la matrice dei regressori di dimensione $n \times p$, y è il vettore dei valori osservati di lunghezza p e W è la matrice dei pesi.
È facile dedurre che la OLS è un caso particolare di WLS quando i pesi sono tutti uguali a 1. Infatti, senza l'applicazione dei pesi la matrice W è una matrice identica I di dimensione $n \times n$ e la soluzione risulta essere $\hat{\beta} = (X^T I X)^{-1} X^T I y = (X^T X)^{-1} X^T y$, ovvero la medesima della regressione ai minimi quadrati ordinari.
Uno dei maggiori vantaggi della WLS la sua efficienza per la gestione degli outliers e per governare al meglio l'eteroschedasticità; riesce inoltre a ricavare informazioni anche da dataset di piccole dimensioni, è capace di gestire situazioni di regressione dove i punti dati sono di qualità variabile e produce stime precise.
Lo svantaggio principale è sicuramente il fatto che questo tipo di modello si basa sull'assunzione che i pesi sono noti; nella maggior parte delle applicazioni reali ciò non è garantito, quindi diventa necessaria una stima degli stessi. Se i pesi sono stimati a partire da un piccolo numero di osservazioni replicate, i risultati dell'analisi possono essere influenzati in modo piuttosto pesante. Ciò è particolarmente probabile quando si calcolano i pesi per valori estremi del predittore o delle variabili esplicative usando solo alcune osservazioni.

3.4 Applicazione del modello WLS

Come riportato precedentemente, al numero di corpuscoli Y è stata applicata una trasformazione logaritmica per assumerne la normalità. Si può quindi affermare che

$$Z = log(Y) \sim N(\mu_Z, \sigma_Z^2)$$

Una variabile aleatoria X il cui logaritmo $log(X)$ si distribuisce normalmente è detta variabile aleatoria log-normale.
Di conseguenza si può notare il numero di corpuscoli Y segue una distribuzione log-normale [22], dato che il suo logaritmo $Z = log(Y)$ segue una distribuzione normale, quindi

$$Y \sim logN(\mu_Y, \sigma_Y^2)$$

È chiaro che nell'ambito di una regressione lineare pesata è di fondamentale importanza la scelta dei pesi w_i.
Un primo metodo per calcolare i pesi è stato porli pari a $w_i = \frac{1}{\sigma_i^2}$ in tal modo hanno più peso le osservazioni con varianza minore; con questa scelta si minimizzano gli standard error delle stime e la varianza delle osservazioni è posta come

$$V(Z_i) = \sigma_i^2 = \frac{1}{w_i}$$

Come già detto, l'intervallo di ampiezza *Ampiezza.Corp=corpuscoli.UL-corpuscoli.LL* può rappresentare una stima della variabilità di *corpuscoli*. Lavorando ad un livello di confidenza $1-\alpha$, si può dire che $P(corpuscoli.LL < Y < corpuscoli.UL) = 1-\alpha$ e quindi
$P(Y < corpuscoli.LL) = \frac{\alpha}{2}$ e $P(Y < corpuscoli.UL) = 1 - \frac{\alpha}{2}$. Essendo i quantili di ordine $\frac{\alpha}{2}$ ed $1 - \frac{\alpha}{2}$ di un intervallo di confidenza delle statistiche d'ordine, sono invarianti rispetto a trasformazioni lineari; è opportuno quindi passare da Y a Z tramite una trasformata logaritmica, ricordando che $Z = log(Y)$.

$$1-\alpha = P(corpuscoli.LL < Y < corpuscoli.UL) = P(log(corpuscoli.LL) < Z < log(corpuscoli.UL))$$

Ne consegue ovviamente che $P(Z < log(corpuscoli.LL)) = \frac{\alpha}{2}$ e che $P(Z < log(corpuscoli.UL)) = 1 - \frac{\alpha}{2}$. Dopo aver costruito un intervallo di confidenza a livello $1-\alpha$ per Z, ovvero $(\mu_Z - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_Z}{\sqrt{n}}, \quad \mu_Z + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_Z}{\sqrt{n}})$, l'ampiezza è pari a

$$ampiezza.IC.log.corpuscoli = (\mu_z + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_Z}{\sqrt{n}}) - (\mu_z - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_Z}{\sqrt{n}}) = 2z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_Z}{\sqrt{n}}$$

Essendo naturalmente $z_{1-\frac{\alpha}{2}}$ il quantile di livello $1 - \frac{\alpha}{2}$ di una normale standard, ovvero 1.96, si può scrivere

$$ampiezza.IC.log.corpuscoli = 2 \cdot 1.96 \cdot \frac{\hat{\sigma}_Z}{\sqrt{n}}$$

ne consegue che per avere gli standard error $\hat{\sigma}_Z$, ovvero la stima della deviazione standard dello stimatore, basta usare la formula inversa

$$\hat{\sigma}_Z = \frac{ampiezza.IC.log.corpuscoli \cdot \sqrt{n}}{2 \cdot 1.96} = \frac{log(corpuscoli.UL - corpuscoli.LL)\sqrt{193}}{3.92}$$

Di conseguenza, dato che $V(Z_i) = \sigma_i^2 = \frac{1}{w_i}$, sono stati ricavati i pesi tramite la formula inversa, ovvero $w_i = \frac{1}{\hat{\sigma}_i^2}$.
È da sottolineare che questa procedura per ricavare $\hat{\sigma}_Z$ è frutto di un'approssimazione,

ritenuta valida alla luce dei dati a disposizione e del caso studio.
Un secondo metodo è stato scegliere i pesi pari del reciproco di un predittore x_i; è noto infatti che se la varianza degli errori è in relazione lineare con un predittore, si può risolvere il problema ponendo $V(\varepsilon_i) = x_i\sigma^2$ ed i pesi uguali a $w_i = \frac{1}{x_i}$. In questo modo si avrà che $V(\varepsilon_i^*) = V(\frac{\varepsilon_i}{\sqrt{x_i}}) = \frac{1}{x_i}V(\varepsilon_i) = \frac{1}{x_i} \cdot x_i\sigma^2 = \sigma^2$ costante $\forall\ i$.
Esistono altri tipi di pesi in casi particolari che non rientrano nel caso di studio, ad esempio se l'$i-$esimo valore è la media di n_i osservazioni si pongono i pesi $w_i = n_i$, oppure se l'$i-$esimo valore è la somma di n_i osservazioni si pongono i pesi $w_i = \frac{1}{n_i}$. Altri possibili pesi che sono stati considerati sono $w_i = \frac{1}{y_i}, w_i = \frac{1}{y_i^2}, w_i = \frac{1}{\hat{y}_i}, w_i = \frac{1}{\hat{y}_i^2}, w_i = \frac{1}{|\sigma_i|}$.
Tra tutti i possibili pesi considerati sono stati scelti $w_i = \frac{1}{\sigma^2}$, in quanto contengono una stima della variabilità delle fibre e sono risultati i migliori in termini di efficienza nell'applicazione del modello.
Sono state candidate le stesse variabili di prima ed anche in questo caso è stato utilizzato un approccio forward e la funzione *add*1 di R; esattamente come prima, è stato tenuto sotto controllo il fenomeno della multicollinearità.
Per valutare la bontà di adattamento del modello si è guardato anche in questo caso alla significatività delle esplicative, gli indici AIC e BIC da minimizzare ed il coefficiente di determinazione $\overline{R^2}$ da massimizzare.

Il modello finale, riassunto nella Tabella 3.4.1 è il medesimo del precedente, vale a dire

$$\log(y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

ed include sempre le variabili *log.fibre*, *anfiboli* e *latenza*.

	Dependent variable:
	y
log.fibre	0.965^{***}
	(0.066)
I(anfiboli <90)	-0.577^{***}
	(0.210)
I(latenza <33)	-0.522^{**}
	(0.227)
Constant	-4.018^{***}
	(0.984)
Observations	193
R^2	0.607
Adjusted R^2	0.600
Akaike Inf. Crit.	691.814
Bayesian Inf. Crit.	708.127
Residual Std. Error	0.561 (df = 189)
F Statistic	97.192^{***} (df = 3; 189)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabella 3.4.1: Modello finale WLS.

L'analisi delle stime dei coefficienti, come si nota dalla Figura 3.4.1, porta a conclusioni pressochè uguali al modello OLS.
La variabile più significativa è sicuramente *log.fibre*; a parità di altre variabili, un aumento dell'1% del numero di fibre corrisponde ad un aumento dello 0.965% di corpuscoli (IC 95%: (0.83, 1.09)). Sempre a parità di altre variabili, la presenza di una percentuale di fibre inferiore al 90% riduce il numero di corpuscoli del 57.7% (IC 95%: (57.27, 58.09)) mentre latenze inferiori a 33 anni riducono la percentuale di corpuscoli del 52.2% (IC 95%: (51.7, 52.6)).

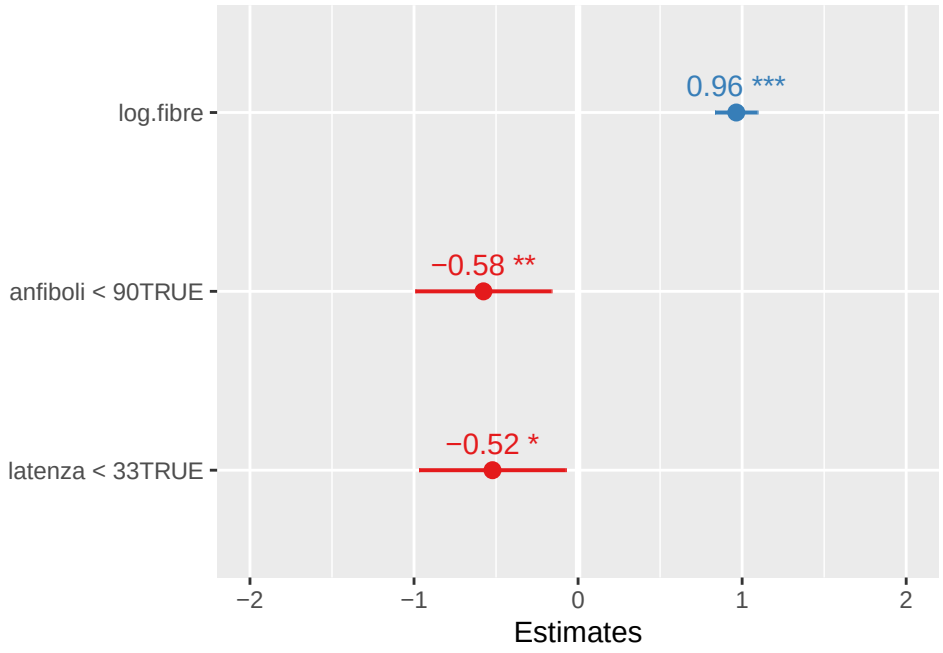

Figura 3.4.1: Rappresentazione grafica delle stime del modello WLS.

Come si nota dalla Tabella 3.4.1, tutte le esplicative sono significative ad un livello del 95%. Per quanto riguarda la bontà di adattamento del modello, si nota che il coefficiente di determinazione R^2 rimane pressochè costante, pari a 0.6067, così come l'$\overline{R^2}$ corretto, pari a 0.6.
Riguardo l'adattamento dei residui, non si nota una grossa differenza con il modello precedente; probabilmente nel modello sono presenti alcuni valori anomali che ne pregiudicano l'adattamento.

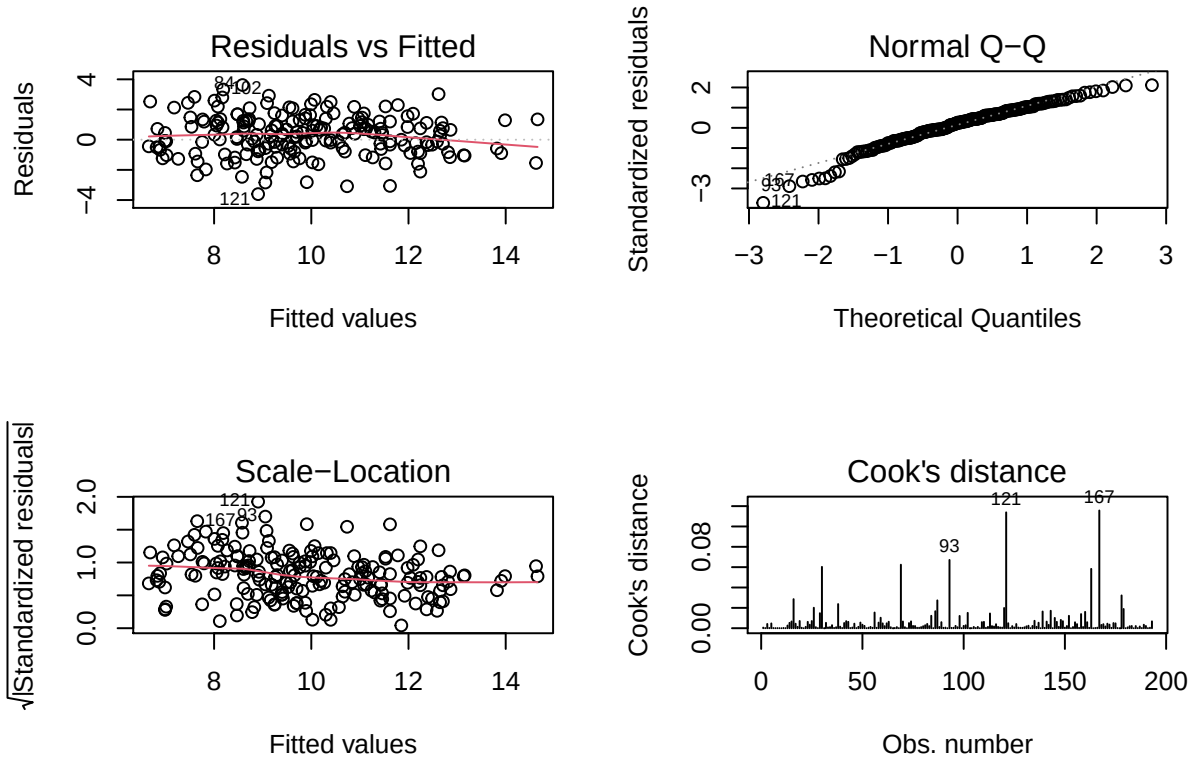

Figura 3.4.2: Adattamento del modello WLS.

4 Studio degli outliers

Come si è visto nel capitolo precedente, il modello finale preso in considerazione è influenzato da valori anomali; è stato quindi dedicato questo capitolo all'approfondimento degli outliers ed alla metodologia per trattarli, al fine di migliorare la bontà del modello finale.

4.1 Introduzione ai valori anomali

In statistica gli outliers (o "valori anomali") [23] sono quei valori che si discostano in maniera significativa dall'andamento generale della distribuzione; essendo estremamente alti o estremamente bassi rispetto all'insieme dei dati, rappresentano casi isolati rispetto al resto della distribuzione.
Non sono da confondere con i valori estremi, che sono definiti come i valori più grandi o più piccoli di una distribuzione, o in generale i valori prossimi alla coda della stessa.
Si dividono generalmente in due gruppi:

- *outlier non rappresentativi*: si tratta di valori anomali causati da errori in fase di compilazione. Un classico caso è costituito dall'utilizzo di un'unità di misura errata (ad esempio euro anzichè migliaia di euro). Sono errori che occorrerebbe individuare e correggere fin da subito;
- *outlier rappresentativi*: si tratta di valori anomali non dovuti ad errori di misurazione, bensì ad eventi relativi all'unità di riferimento non (del tutto) valutabili sulla base delle informazioni disponibili su di essa.

Altre cause possono essere delle circostanze eccezionali che hanno influenzato la misurazione e l'osservazione del fenomeno, oppure la contaminazione di un insieme di valori che proviene da una diversa realtà rispetto a quella della maggior parte dei dati. Nel primo caso, è opportuno eliminare le osservazioni anomale per evitare che influenzino le analisi a posteriori, nel secondo caso conviene svolgere un'analisi specifica.
Esistono diversi metodi per catalogare gli outliers, uno di questi è quello del range interquantile (in inglese IQR, "interquartile range").
Un valore x^* è detto valore anomalo se

$$|x^* - Q_1| > 1.5(Q_3 - Q_1) \quad \textit{oppure} \quad |x^* - Q_3| > 1.5(Q_3 - Q_1)$$

ed è detto fortemente anomalo se

$$|x^* - Q_1| > 3(Q_3 - Q_1) \quad \textit{oppure} \quad |x^* - Q_3| > 3(Q_3 - Q_1)$$

dove Q_1 e Q_3 sono il primo ed il terzo quartile e la loro differenza, cioè $Q_3 - Q_1$ è detta differenza o range interquantile.
I valori anomali possono influenzare svariati indicatori, come la media aritmetica o la varianza; possono inoltre condizionare anche gli indici di associazione tra variabili come il coefficiente di correlazione di Pearson.

È bene tenere a mente che per analizzare un campione contenente outliers ci sono indici che sono più sensibili come la media aritmetica ed altri che ne risentono meno come la mediana o la media geometrica; un'altra scelta è la media troncata, ottenuta eliminando il 5% dei valori più alti o più bassi.
Una strategia può essere quella di eliminare dal dataset i valori anomali per proseguire con un'analisi "pulita", ma è un metodo non senza rischi che occorre valutare accuratamente per evitare errori metodologici.

4.2 Metodo di Farcomeni-Viviani

Come già visto in precedenza, anche i risultati di un modello di regressione possono essere pesantemente influenzati dagli outliers, che ne compromettono bontà di adattamento e qualità dei residui.
A questo proposito lo studio di Farcomeni e Viviani (2011) [24] mira a correggere queste distorsioni con un approccio basato sul trimming (in italiano "rifinitura"), una tecnica statistica che consiste in un processo di rimozione o esclusione di valori estremi o valori anomali da un dataset.
L'algoritmo si basa su un ciclo di procedure accetto-rifiuto che consente di eliminare le unità statistiche che peggiorano la log-verosimiglianza del modello. Partendo dal presupposto che il modello finale conterrà meno osservazioni del modello originale, visto che verranno eliminate alcune unità statistiche, ad ogni iterazione viene candidato un valore ad entrare nel modello al posto di un altro e viene calcolata la massima log-verosimiglianza parziale ottenuta.
L'algoritmo viene inizializzato da un insieme I di osservazioni di cardinalità $n \cdot (1 - \alpha)$; a seconda del problema preso in considerazione lo si può scegliere manualmente considerando i valori non anomali, oppure può essere scelto casualmente.
Il valore α è fissato ed indica la percentuale di valori anomali, è quindi compreso tra 0 e 1; in questo caso è stato scelto $\alpha = 0.05$, ma può anche essere pari a 0.02 o 0.1, a seconda dello studio.
Un parametro importante da inserire è il numero di iterazioni k_{max}, qui posto pari a 10000. L'insieme $I^{(k)}$ è composto dalle osservazioni presenti nel modello alla k-esima iterazione, mentre l'insieme $C(I^{(k)})$ è il suo complementare.
Il secondo parametro da inserire è D ed è relativo alla massima variazione attesa nella log-verosimiglianza parziale quando cambia un elemento di I. La sua scelta influisce solo sulla velocità della convergenza dell'algoritmo e deve essere grande a sufficienza per assicurarla; è stato posto pari a $0.1 \cdot n \cdot (1 - \alpha)$.
Durante ognuna delle k_{max} iterazioni vengono campionati due valori $i \in I^{(k)}$ ed $i' \in C(I^{(k)})$ e viene calcolata la massima log-verosimiglianza parziale del modello corrente calcolato con i valori di $I^{(k)}$ e quella del modello "candidato" $I^{(k)}_{cand}$ calcolato sostituendo i con i' in $I^{(k)}$. Ogni volta che questo massimo è maggiore della massima log-verosimiglianza calcolata nel sottoinsieme $I^{(k)}$ viene accettato con probabilità 1. Se invece la log-verosimiglianza non

aumenta, viene accettato con probabilità $p < 1$.
Viene qui sotto riportato l'algoritmo di Farcomeni-Viviani.

for $k = 1, ..., k_{max}$ **do**

campiono un candidato $i' \in C(I^{(k)})$ ed un candidato $i \in I^{(k)}$ con

probabilità uniforme

Introduco $I_{cand} = (I^{(k)} \setminus i) \cup i'$, ottenuto sostituendo i con i' in $I^{(k)}$.

Pongo $\tau_k = log(k+1)/D$ e $p = min(e^{\tau_k(log(L(\hat{\beta}_{cand}, I_{cand})) - log(L(\hat{\beta}, I^{(k)})))}, 1)$,

dati $\hat{\beta} := \underset{\beta}{argmax}\ L(\beta, I^{(k)})$ e $\hat{\beta}_{cand} := \underset{\beta}{argmax}\ L(\beta, I_{cand})$.

Introduco U estrazione casuale da una variabile casuale di Bernoulli con

parametro p.

if $U = 1$ **then**

$I^{(k+1)} = I_{cand}$.

else

$I^{(k+1)} = I^{(k)}$.

end if

end for

4.3 Applicazione nel modello e risultati

L'algoritmo Farcomeni-Viviani è stato quindi applicato al modello di regressione pesata finale, che nonostante un lieve miglioramento rispetto al modello di regressione classico, presenta problemi in termini di outliers. La procedura impiega circa 16 secondi a convergere. L'interpretazione dei coefficienti di regressione stimati è simile al Capitolo 3, come si vede dalla Tabella 4.3.1 e dalla Figura 4.3.1. A parità di altre variabili, un aumento dell'1% del numero di fibre corrisponde ad un aumento dello 0.932% di corpuscoli (IC 95%: (0.82, 1.04)). Sempre a parità di altre variabili, la presenza di una percentuale di fibre inferiore al 90% riduce il numero di corpuscoli del 49.8% (IC 95%: (49.5, 50.21)) mentre latenze

inferiori a 33 anni riducono la percentuale di corpuscoli del 65.6% (IC 95%: (65.2, 65.95)). Riguardo la bontà di adattamento del modello, si nota un R^2 discretamente alto (0.6925) ed anche la sua versione corretta (0.6874).

	Dependent variable:
	y
log.fibre	0.934***
	(0.055)
I(anfiboli <90)	−0.499***
	(0.180)
I(latenza <33)	−0.653***
	(0.192)
Constant	−3.349***
	(0.826)
Observations	183
R^2	0.693
Adjusted R^2	0.688
Akaike Inf. Crit.	585.256
Bayesian Inf. Crit.	601.303
Residual Std. Error	0.455 (df = 179)
F Statistic	134.527*** (df = 3; 179)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabella 4.3.1: Modello finale trimmed.

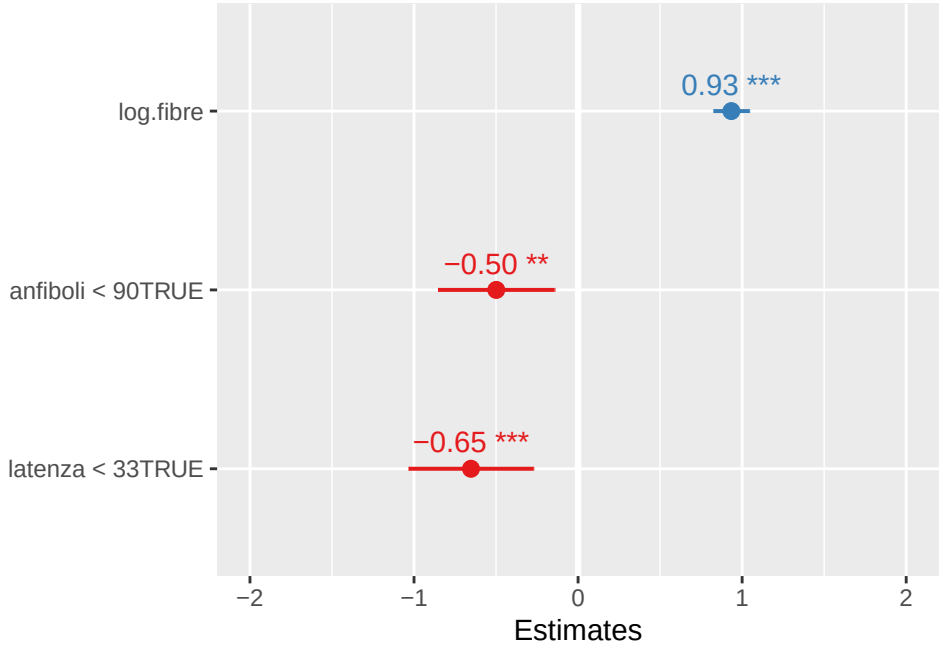

Figura 4.3.1: Rappresentazione grafica delle stime del modello trimmed.

Grazie alla procedura del trimming sono state tolte 10 osservazioni dal modello iniziale (circa il 5% di 193): nella tabella sottostante vengono riportate le loro caratteristiche.

id	corpuscoli	fibre	anfiboli	latenza	w.i
16	5200	1.1e+07	100	48.5	0.2055
30	1200	3200000	100	27.5	0.2397
69	350	675277.7	67	30.5	0.3581
84	2e+05	860000	83	41	0.1059
93	500	1400000	89	33	0.3146
121	200	1200000	21	44.5	0.322
143	2100	4400000	100	36	0.1854
163	450	850000	89	49	0.3321
167	200	180000	100	48.5	0.3755
178	1000	790000	100	35	0.3133

Tabella 4.3.2: Caratteristiche delle osservazioni eliminate.

Anche i risultati derivanti dai residui sono molto buoni, dal momento che sono state eliminate alcune osservazioni anomale. Per essere sicuri di questi risultati sono stati applicati dei test, oltre alla semplice verifica grafica.
Il primo è il test di Breusch-Pagan (1979) [25] e serve a verificare l'ipotesi di omoschedasticità in un modello di regressione lineare; il grafico di riferimento è quello in alto a sinistra della Figura 4.3.2, che confronta i valori previsti con i residui del modello. In situazioni di omoschedasticità il grafico ha un andamento casuale e non cambia la variabilità dei residui a seconda dei valori previsti. Il test è basato su un modello del tipo $\sigma_i^2 = h(z_i^T\gamma)$ dove $z_i = (1, z_{2i}, ..., z_{pi})$ definiscono le differenze tra le varianze; saggiare il test d'ipotesi di omoschedasticità dei residui equivale a verificare il sistema d'ipotesi

$$\begin{cases} H_0 : \gamma_2 = ... = \gamma_p = 0 \\ H_1 : \exists!(\gamma_i, \gamma_j) | \gamma_i \neq \gamma_j, \quad i \neq j \end{cases}$$

Per costruire la statistica test, per prima cosa si applica il modello

$$Y_i = X\beta_i + \varepsilon_i \qquad i = 1, ..., n$$

poi si calcolano dei residui chiamati g_i, tramite la formula $g_i = \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2}$, dove $\hat{\varepsilon}_i$ sono i residui stimati dal modello e $\hat{\sigma}^2$ è la stima di massima verosimiglianza della varianza dell'errore, ovvero $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n}$. Successivamente si imputa un secondo modello

$$g_i = \gamma_1 + \gamma_2 z_{2i} + ... + \gamma_p z_{pi} + \eta_i$$

dove z_i sono solitamente uguali alle covariate originali x_i. La statistica test è calcolata come $LM = 0.5(TSS - SSR)$ dove TSS è la somma degli scarti al quadrato di g_i dalla loro media unitaria e SSR è la somma degli scarti al quadrato della seconda regressione. Sotto

H_0 si distribuisce come una χ^2_{p-1}.
Un secondo test è il test di normalità di Shapiro-Wilk sui residui studentizzati (chiamati talvolta "standardizzati") [26] nel grafico in alto a sinistra della Figura 4.3.2, che confronta i residui studentizzati con i residui teorici in caso di normalità.
In un modello di regressione ai minimi quadrati ordinari sono calcolati come:

$$r_i = \frac{y_i - \hat{y}}{\sqrt{S^2(1-h_{ii})}}$$

dove y_i sono i valori della variabile risposta, $\hat{y}$ quelli previsti dal modello, h_{ii} è l'i-esimo elemento sulla diagonale principale della matrice di proiezione $H = X(X^TX)^{-1}X^T$, pari ad $x_i(X^TX)^{-1}x_i^T$, mentre S^2 è detto "residual standard error", è uno stimatore corretto di σ^2 ed è pari a $\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-p}$ (n è il numero di osservazioni e p il numero di esplicative).
Nel caso trattato in questo capitolo è stato però usato anche un modello di regressione ai minimi quadrati pesati, di conseguenze i residui studentizzati sono

$$r_i = \frac{\sqrt{w_i}\cdot(y_i - \hat{y})}{\sqrt{S^2(1-h_{ii})}}$$

dove w_i sono chiaramente i pesi adottati nel modello ed $S^2 = \frac{\sum_{i=1}^{n} w_i\cdot(y_i-\hat{y})^2}{n-p}$. In entrambi i casi la loro distribuzione è una normale standard.
Con un p-value di 0.318 si accetta l'ipotesi nulla di normalità dei residui studentizzati.

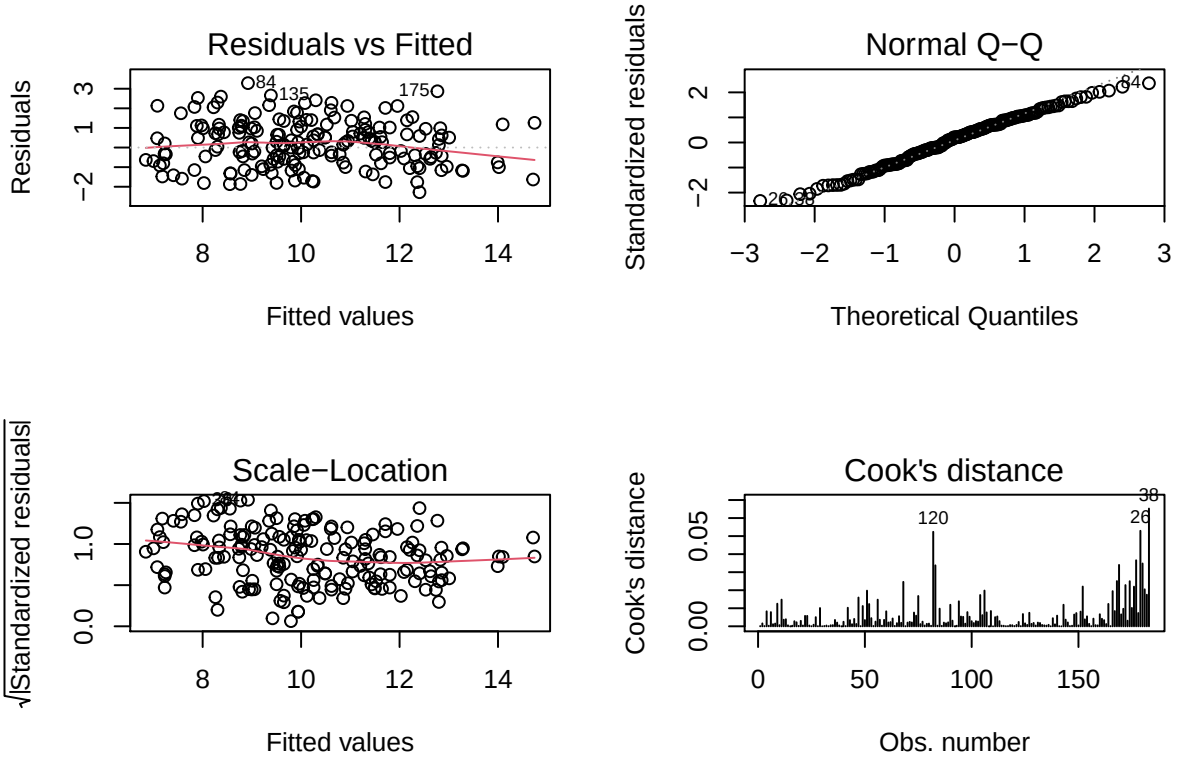

Figura 4.3.2: Adattamento del modello trimmed.

Il confronto tra il modello finale WLS e quello ottenuto grazie al trimming è del tutto a favore di quest'ultimo. Oltre ad un $\overline{R^2}$ nettamente più alto (0.6925 contro 0.6067), presenta anche gli indici AIC e BIC più bassi, rispettivamente 585.2 contro 691.8 e 601.3 contro 708.1.
Una verifica interessante è anche quella sulla bontà dei residui (in alto il modello WLS, in basso il modello trimmed). Il grafico di dispersione non sembra cambiare molto, mentre la normalità dei residui sembra nettamente migliorata nel modello trimmed.

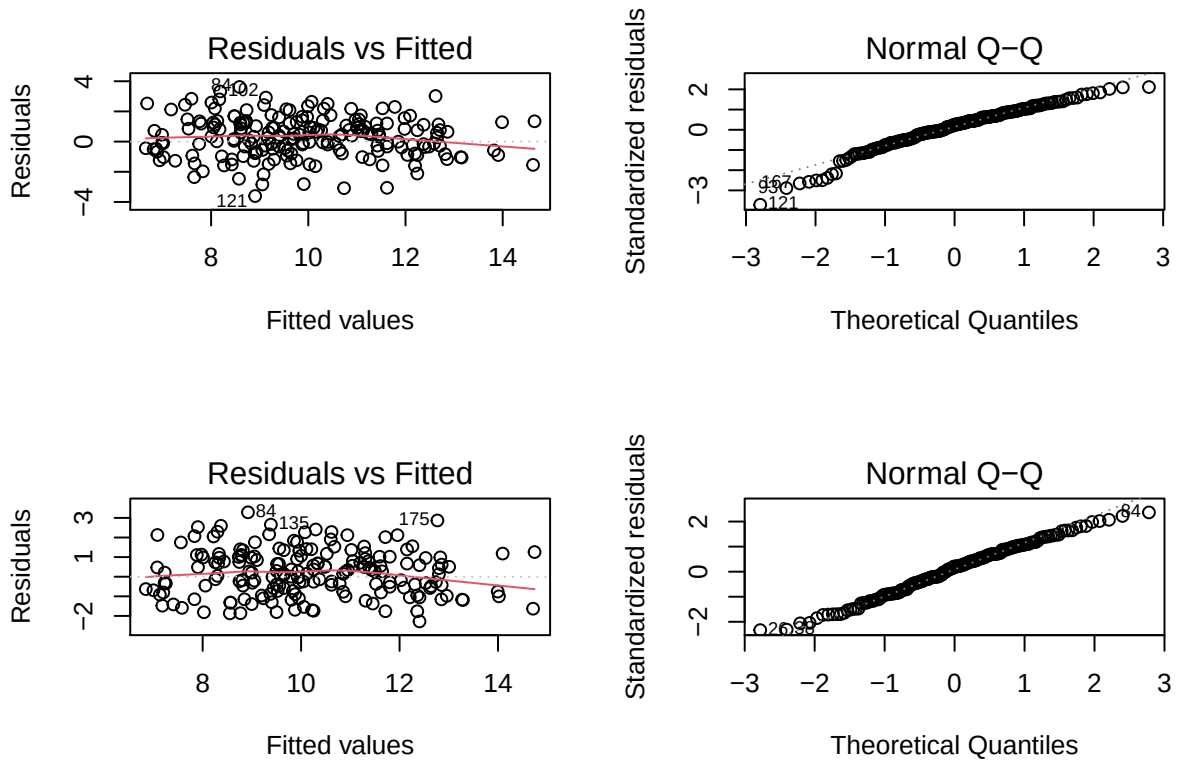

Figura 4.3.4: Confronto grafico tra modello WLS e modello trimmed.

5 Conclusioni

La stesura di questo elaborato ha permesso di sintetizzare ad analizzare approfonditamente le informazioni contenute del dataset oggetto di studio, estremamente ricco e mai trattato precedentemente.
Dopo una lunga ed accurata operazione di pulizia del dataset, in cui si sono organizzate meglio le informazioni e si è risolto il problema dei dati mancanti, le analisi univariate hanno dato una preliminare immagine delle variabili in gioco. Le analisi bivarate hanno invece dato un'idea marginale della relazione tra esplicative e quantità d'interesse, ovvero il numero di corpuscoli di amianto relativo ad ogni soggetto.
Durante queste analisi è emerso che i fattori più rilevanti sono il genere, la regione di residenza, il settore lavorativo, la causa dell'esposizione, la causa ed il livello dell'esposizione. Si è infatti trovato un numero maggiore di corpuscoli negli uomini, residenti in Friuli e con un alto livello di esposizione e che sono entrati a contatto con l'amianto lavorando nei cantieri navali.
Questi risultati sono chiaramente delle osservazioni marginali e per un'analisi più approfondita sono stati utilizzati dei modelli di regressione. Prima un modello di regressione OLS, successivamente un modello WLS ed infine un modello "trimmed" per il trattamento dei valori anomali. In tutti i modelli è emerso che le quantità più rilevanti nel processo di formazione dei corpuscoli sono il numero di fibre, la percentuale di anfiboli sul totale delle fibre ed il periodo di latenza.
In particolare un alto numero di corpuscoli è associato ad un grande numero di fibre, specialmente di anfiboli e ad una latenza lunga, superiore ai 33 anni.
I risultati ottenuti da questo lavoro sembrano essere piuttosto in linea con la letteratura scientifica riguardo l'argomento e possono dare un contributo alla ricerca dei fattori che influenzano la formazione di corpuscoli di amianto.

Appendice A. Output dei test

Test statistic	P value
0.9785	0.04054 *

Tabella A.1: Test di Shapiro-Wilk del numero di corpuscoli per *malattia*=Mesotelioma (scala logaritmica).

Test statistic	P value
0.9781	0.4126

Tabella A.2: Test di Shapiro-Wilk del numero di corpuscoli per *malattia*=Tumore al polmone (scala logaritmica).

Test statistic	P value
0.9384	0.5022

Tabella A.3: Test di Shapiro-Wilk del numero di corpuscoli per *malattia*=Altro (scala logaritmica).

Test statistic	P value
0.3381	0.7136

Tabella A.4: Test di Levene del numero di corpuscoli per la malattia (scala logaritmica).

Test statistic	P value
0.9853	0.3291

Tabella A.5: Test di Shapiro-Wilk del numero di corpuscoli per *regione*=Friuli (scala logaritmica).

Test statistic	P value
0.9613	0.07397

Tabella A.6: Test di Shapiro-Wilk del numero di corpuscoli per *regione*=Veneto (scala logaritmica).

Test statistic	P value
0.985	0.8881

Tabella A.7: Test di Shapiro-Wilk del numero di corpuscoli per *regione*=Altro (scala logaritmica).

Test statistic	P value
0.2357	0.7902

Tabella A.8: Test di Levene del numero di corpuscoli per la regione di residenza (scala logaritmica).

Test statistic	P value
0.9793	0.09923

Tabella A.9: Test di Shapiro-Wilk del numero di corpuscoli per *fonte*=Gorizia (scala logaritmica).

Test statistic	P value
0.9685	0.1583

Tabella A.10: Test di Shapiro-Wilk del numero di corpuscoli per *fonte*=Padova (scala logaritmica).

Test statistic	P value
0.9833	0.8776

Tabella A.11: Test di Shapiro-Wilk del numero di corpuscoli per *fonte*=Altro (scala logaritmica).

Test statistic	P value
0.2238	0.7997

Tabella A.12: Test di Levene del numero di corpuscoli per la fonte (scala logaritmica).

Test statistic	P value
0.9875	0.3722

Tabella A.13: Test di Shapiro-Wilk del numero di corpuscoli per *settore.lavorativo*=Cantieri navali (scala logaritmica).

Test statistic	P value
0.8776	0.03558 *

Tabella A.14: Test di Shapiro-Wilk del numero di corpuscoli per *settore.lavorativo*=Ferrovie (scala logaritmica).

Test statistic	P value
0.9113	0.2524

Tabella A.15: Test di Shapiro-Wilk del numero di corpuscoli per *settore.lavorativo*=Industria (scala logaritmica).

Test statistic	P value
0.976	0.3868

Tabella A.16: Test di Shapiro-Wilk del numero di corpuscoli per *settore.lavorativo*=Altro (scala logaritmica).

Test statistic	P value
0.8674	0.459

Tabella A.17: Test di Levene del numero di corpuscoli per la causa d'esposizione (scala logaritmica).

Test statistic	P value
0.9554	0.2052

Tabella A.18: Test di Shapiro-Wilk del numero di corpuscoli per *livello.esposizione*=Alto (scala logaritmica).

Test statistic	P value
0.9689	0.04864 *

Tabella A.19: Test di Shapiro-Wilk del numero di corpuscoli per *livello.esposizione*=Medio (scala logaritmica).

Test statistic	P value
0.9689	0.4322

Tabella A.20: Test di Shapiro-Wilk del numero di corpuscoli per *livello.esposizione*=Basso (scala logaritmica).

Test statistic	P value
0.9785	0.5304

Tabella A.21: Test di Shapiro-Wilk del numero di corpuscoli per *livello.esposizione*=Altro (scala logaritmica).

Test statistic	P value
0.1545	0.9267

Tabella A.22: Test di Levene del numero di corpuscoli per il livello d'esposizione (scala logaritmica).

Test statistic	df	P value
170.9	4	6.789e-36 * * *

Tabella A.23: Test χ^2 di Pearson per la regione e la fonte.

Test statistic	df	P value
119	6	2.587e-23 * * *

Tabella A.24: Test χ^2 di Pearson per la fonte ed il settore lavorativo.

Test statistic	df	P value
3.614	3	0.3062

Tabella A.25: Test di Breusch-Pagan per il modello trimmed.

Test statistic	P value
0.9911	0.318

Tabella A.26: Test di Shapiro-Wilk per i residui studentizzati del modello trimmed.

Appendice B. Codice dell'algoritmo di Farcomeni-Viviani

```
db<-data.frame(y, log.fibre, anfiboli, latenza,w.i)
M=lm(y~1+log.fibre+I(anfiboli<90)+I(latenza<33),weights = w.i)
attach(db)
n<-dim(db)[1]
k<-10000 #Numero massimo di iterazioni
alpha<-0.05 #Fisso il 5% di valori anomali
D<-0.01*n*(1-alpha)
n_sample<-n*(1-alpha) #Numerosità del dataset trimmed

I_iniziale<-sample(n,n_sample,replace=FALSE) #Dataset trimmed iniziale
C_I_iniziale<-(1:n)[!( 1:n%in% I_iniziale)]  #Dataset degli esclusi
##Modello iniziale
attach(db)
fit_k<-lm(formula=M, data=db[I_iniziale,])
I_finale<-NULL #Dataset finale, imposto un valore nullo


I_nuovo<-I_iniziale #Dataset trimmed nuovo (all'inizio imposto uno a caso)
C_I_nuovo<-C_I_iniziale #Dataset degli esclusi iniziale

#Inizio il ciclo
for(i in 1:k){
I<-sample(I_nuovo,1,replace=FALSE) #Nuovo candidato ad essere sostituito
Iprimo<-sample(C_I_nuovo,1,replace=FALSE) #Candidato da sostituire
I_c<-c(I_nuovo[!(I_nuovo %in% I)],Iprimo) #Rimpiazzo, creo nuovo dataset trimmed
fit_c<-lm(formula = M ,data=db[I_c,]) #Effettuo il fit
tau_k=log(i+1)/D #Calcolo la costante tau_k
p<-min(exp(tau_k*(logLik(fit_c)-logLik(fit_k))),1)
#Calcolo la probabilità p simile ad accetto/rifiuto
#Se il risultato è 1 utilizzo il nuovo dataset, altrimenti proseguo
if (rbinom(1,1,p)) {
  I_nuovo<-I_c
  C_I_nuovo<-(1:n)[!( 1:n%in% I_c)]
  I_finale<-I_c
  fit_k<- fit_c
}
I_finale
```

```
}

#Come risultato nell'oggetto I_finale ci sono le osservazioni che tengo
Mf<-lm(formula = M, data=db[I_finale,]) #Nuovo fit trimmed
summary(Mf)
#Caratteristiche dei soggetti esclusi
db[-I_finale,]
```

Bibliografia

[1] Harri Vainio, Panu Oksa, Timo Tuomi, Tapio Vehmas, Henrik Wolff1, *Aggiornamento dei Criteri di Helsinki 2014*, epiprev.it.

[2] John T. Hodgson and Andrew Darnton (2000), *The Quantitative Risks of Mesothelioma and Lung Cancer in Relation to Asbestos Exposure*, Epidemiology and Medical Statistics Unit, Health and Safety Executive, Magdalen House, Stanley Precinct, Bootle L20 3QZ, UK.

[3] Andrew M. Churg, and Martha L. Warnock (1981), *Asbestos and Other Ferruginous Bodies. Their Formation and Clinical Significance*, Department of Pathology, University of California, San Francisco, California.

[4] Anna Somigliana, Paolo Girardi, Pietro G. Barbieri, Enzo Merler (2016), *Factors influencing the asbestos bodies among pleural mesotheliomas and lung cancers examined for retained asbestos fibres*, 13th International Conference of the International Mesothelioma Interest Group San Francisco, California.

[5] C. Paris, F. Galateau-Salle, C. Creveuil, R. Morello, C. Raffaelliz, J.C. Gillon, M.A. Billon-Gallandf, J.C. Pairon, L. Chevreau, M. Letourneux (2002), *Asbestos bodies in the sputum of asbestos workers: correlation with occupational exposure*, European Respiratory Journal.

[6] Pratan Vathesatogkit, Timothy J. Harkin, Doreen J. Addrizzo-Harris, Marion Bodkin, Michael Crane, William N. Rom(2004), *Clinical Correlation of Asbestos Bodies in BAL Fluid*, Occupational and environmental lung disease.

[7] *Sintesi delle conoscenze relative all'esposizione e al profilo tossicologico: Amianto*, Ministero della Salute.

[8] Ahmed Naeem, Sachchida N. Rai, Louisdon Pierre (2020) *Histology, Alveolar Macrophages*, StatPearls.

[9] *What can sputum tell us?*, MedicalNewsToday.

[10] *Amianto - Dove è stato utilizzato*, ARPAT.

[11] Agenzia Zoe *Mesotelioma*, AIRC.

[12] *Asbestosis*, asbestos.com.

[13] *Corpuscoli dell'asbesto nel tessuto polmonare umano e liquidi biologici: metodo analitico e atlante fotografico*, ISS.

[14] *Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi*, ISTAT.

[15] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, Philip J. Leaf (2011), *Multiple imputation by chained equations: what is it and how does it work?*, NCBI.

[16] Stef van Buuren, Karin Groothuis-Oudshoorn (2011), *mice: Multivariate Imputation by Chained Equations in R*, Journal of Statistical Software.

[17] Matteo Grigoletto, Francesco Pauli, Laura Ventura (2017), *MODELLO LINEARE. Teoria e applicazioni con R*, Università degli Studi di Padova.

[18] Roderick J. A. Little, Donald B. Rubin, *Statistical analysis with Missing Data*, Univerisity of California at Los Angeles, Harvard University.

[19] Laura Ventura, Walter Racugno (2017), *Biostatistica. Casi di studio in R*, Università degli Studi di Padova, Università degli Studi di Cagliari.

[20] Daniel Lüdecke (2020), *Plotting Estimates (Fixed Effects) of Regression Models*, Comprehensive R Archive Network.

[21] *Nonconstant Variance and Weighted Least Squares*, PennState Elbery College of Science.

[22] Sara Gustavsson, Björn Fagerberg, Gerd Sallsten,Eva M. Andersson, *Regression Models for Log-Normal Data: Comparing Different Methods for Quantifying the Association between Abdominal Adiposity and Biomarkers of Inflammation and Insulin Resistance*, NCBI.

[23] Roberto Gismondi, *Metodi per il trattamento dei dati anomali nelleindagini longitudinali finalizzate alla stima di variazioni*, ISTAT.

[24] Alessio Farcomeni, Sara Viviani, *Robust estimation for the Cox regression model based on trimming*, Biometrical Journal.

[25] Roberto Pedace, *The Role of the Breusch-Pagan Test in Econometrics*, Dummies.

[26] *Introduction to Regression Analysis*, NC State University.