Università
Ca'Foscari
Venezia

Corso di Laurea magistrale
in Statistica Per l'Impresa

Tesi di Laurea

Il fenomeno del fumo in Italia: un'analisi statistica dell'evoluzione temporale attraverso modelli lineari generalizzati e modelli multilivello.

Relatore
prof. Andrea Pastore

Laureanda
Laura Pezzato
Matricola 815796

Anno Accademico
2011-2012

Indice

Elenco delle tabelle

Elenco delle figure

Introduzione

Il fenomeno del fumo in Italia viene esaminato tramite un'analisi di serie storiche improntata sui modelli di regressione. Ci poniamo l'obiettivo di individuare, a livello globale ed all'interno di determinati gruppi di unità statistiche, una componente di trend ed un'eventuale componente stagionale; tuttavia, sulla presenza di quest'ultima ci permettiamo di mantenere delle riserve.

Nel primo capitolo dell'elaborato presentiamo un'analisi descrittiva dei dati, al fine di avere una panoramica iniziale della loro struttura.

Nel secondo capitolo esaminiamo le serie storiche implementando modelli lineari generalizzati basati sulla distribuzione Beta-Binomiale e considerando diverse sottopopolazioni, definite sulla base di alcune variabili socio-demografiche (classe d'età, sesso, livello d'istruzione, regione d'appartenenza). Questo approccio ci consente di individuare eventuali componenti di trend e di stagionalità.

Il terzo capitolo è dedicato ai modelli multilivello, adatti a rappresentare e stimare fenomeni sia a livello globale sia a livello di sottopopolazioni, permettendo di cogliere ed interpretare differenze ed analogie tra gruppi di unità statistiche.

Capitolo 1

Analisi descrittiva

1.1 Presentazione del problema

I dati, cui facciamo riferimento nelle analisi seguenti, provengono dalla rilevazione sistematica e continua delle abitudini, degli stili di vita e dello stato di attuazione dei programmi di intervento che l'Italia sta realizzando per modificare i comportamenti a rischio.

Il sistema di sorveglianza Passi [Passi], attraverso questa raccolta di dati, si prefigge l'obiettivo di effettuare un monitoraggio il più possibile completo sullo stato di salute della popolazione adulta italiana. In particolare, il questionario (consultabile a pagina 152), oltre alla rilevazione di alcuni dati socio-anagrafici, presenta domande riguardanti lo stato di salute, la qualità della vita percepita, l'attività fisica svolta, l'abitudine al fumo, l'alimentazione, il rapporto con l'alcol, la sicurezza stradale ed una serie di quesiti su esami e farmaci specifici di alcune patologie.

L'interpretazione dei dati è messa a disposizione di coloro che devono progettare, realizzare e valutare interventi in salute pubblica.

La popolazione di studio è costituita dalle persone di 18-69 anni iscritte nelle liste delle anagrafi sanitarie delle Aziende Sanitarie Locali partecipanti a Passi. Il campionamento si fonda su un campione mensile stratificato proporzionale, per

sesso ed età, la cui dimensione minima per ciascuna Asl è di 25 unità. Tutte le regioni italiane hanno aderito al sistema di sorveglianza Passi.

Ogni unità selezionata è intervistata telefonicamente, previo avviso tramite lettera da parte della relativa Asl; l'intervista prevede la somministrazione di un questionario standardizzato e validato a livello nazionale ed internazionale.

Il dataset di interesse è costituito da 172162 unità e contiene i dati raccolti tramite le interviste effettuate mensilmente da febbraio 2007 a dicembre 2011; è d'obbligo precisare che luglio ed agosto sono stati considerati come un'unica mensilità durante la fase di raccolta dei dati, mentre per le nostre analisi ci riferiamo a luglio e ad agosto come a due mensilità, sdoppiando i pattern di risposta ottenuti .

Alcune unità sono state scartate dall'analisi: per i mesi di febbraio 2007 e marzo 2007 è stata rilevata una sola osservazione, dunque le misure di sintesi di quei periodi sarebbero risultate distorte a causa dell'anomala numerosità ridotta; alcune osservazioni mancano dell'indicazione del periodo cui facevano riferimento e non sarebbe stato possibile includerle correttamente nell'analisi di serie storiche.

Per le successive analisi faremo particolare riferimento alle sezioni 3 (abitudine al fumo) e 14 (dati socio-anagrafici) del questionario: molte variabili di interesse riguardano le domande contenute in esse, altre sono variabili spazio-temporali.

Cercheremo di valutare l'evoluzione temporale del fenomeno del fumo ossia di individuare una tendenza di fondo ed un'eventuale componente stagionale, che faccia riferimento ad un andamento periodico del fenomeno oppure a degli effetti di mese. Mentre ci aspettiamo con relativa certezza di rilevare una componente di trend, non siamo altrettanto convinti di riscontrare la presenza di stagionalità, poichè siamo di fronte ad un fenomeno a cui intuitivamente non riusciamo ad accostare un andamento che presenti movimenti ciclici o regolarità

mensili.

L'analisi verrà svolta considerando le unità statistiche dapprima globalmente e poi partizionate secondo determinati criteri di stratificazione, poichè supponiamo che vi siano differenze significative tra gli andamenti del fenomeno riscontrati all'interno di specifiche sottopopolazioni. Queste ultime verranno identificate sulla base di quattro caratteristiche: classe d'età, sesso, livello d'istruzione, regione di appartenenza dell'intervistato.

1.2 Esplorazione dei dati

1.2.1 La variabile `fumatore`

Iniziamo ad analizzare i dati concentrandoci sulla variabile dicotomica `fumatore`, che vale 1 se l'individuo è fumatore e 0 se non lo è, ed andando ad estrarre delle serie storiche che rappresentino l'andamento di questa variabile all'interno di determinati gruppi di unità, individuati via via tramite diverse variabili di stratificazione.

Andremo poi a verificare la presenza o meno di una componente stagionale all'interno delle serie considerate, attraverso il metodo delle medie mobili e l'applicazione dell'operatore differenza [Ricci 2005]. Il valore y_t al tempo t della variabile d'interesse può essere espresso come

$$y_t = T_t + S_t + \varepsilon_t$$

dove T_t è la componente di trend, S_t è la componente stagionale, ε_t è la componente erratica; la componente stagionale può essere stimata come differenza tra il valore y_t e la stima della componente di trend. Quest'ultima viene calcolata con un'opportuna ponderazione dei valori della serie:

$$T_t = \frac{1}{2a+1} \sum_{j=-a}^{a} X_{t-j};$$

nel caso di osservazioni con ricorrenza mensile, $a = 6$.

Riportiamo il numero di osservazioni raccolte in ogni mese (tabella 1.1), facendo

attenzione al fatto che i primi due mesi hanno numerosità nettamente minori rispetto ai successivi.

Tabella 1.1: Numero di osservazioni rilevate in ogni mese

Mese	Numerosità
200704	987
200705	1582
200706	2785
200707	3320
200708	3320
200709	3237
200710	3491
200711	3259
200712	3131
200801	3237
200802	3408
200803	3364
200804	3374
200805	3407
200806	3237
200807	3659
200808	3659
200809	3422
200810	3819
200811	3425
200812	3274
200901	3407
200902	3443
200903	3449
200904	3409
200905	4180
200906	3925
200907	3999
200908	3999

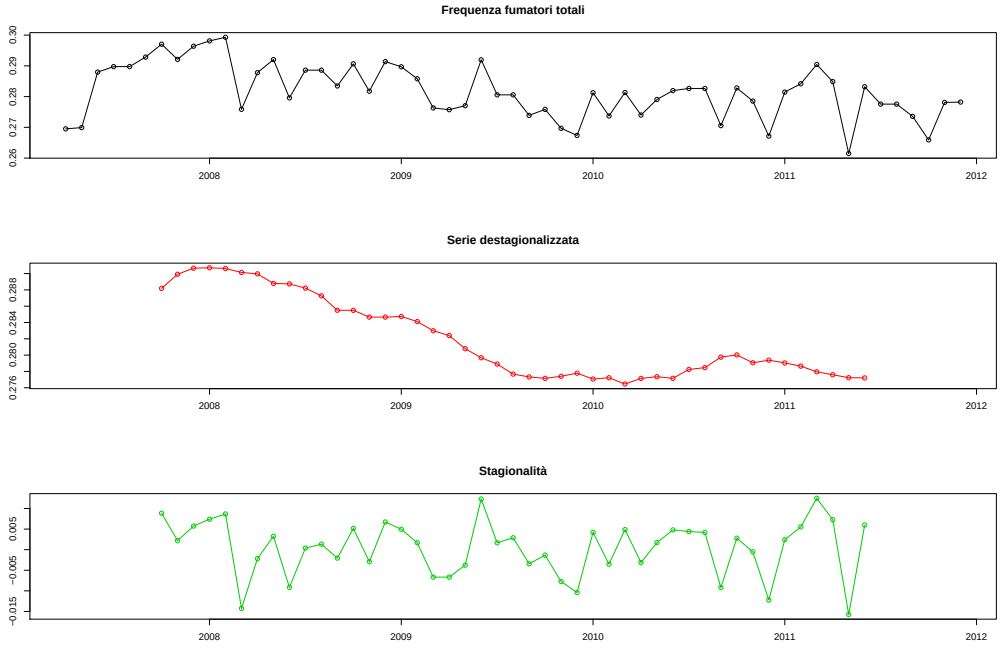
Mese	Numerosità
200909	3421
200910	3426
200911	3315
200912	3239
201001	3015
201002	3058
201003	3096
201004	3281
201005	3268
201006	3419
201007	3913
201008	3913
201009	3308
201010	3398
201011	3285
201012	3309
201101	3297
201102	3255
201103	3450
201104	3419
201105	3419
201106	3358
201107	3675
201108	3675
201109	3469
201110	3253
201111	3301
201112	3260

1.2.1.1 Andamento globale

L'andamento della percentuale media di fumatori registrati in ogni mese del periodo di osservazione considerato è rappresentato nel grafico 1.1, che include anche l'andamento del trend stimato tramite il metodo delle medie mobili e la

componente stagionale calcolata per differenza tra le due serie di cui sopra. La media globale è circa del 28.28%, la variabilità attorno a questo valore è contenuta: vi sono punti minimi che superano di poco il 26% e picchi che sfiorano il 30%. Si può notare un trend tendenzialmente decrescente, ossia sembra che la percentuale di fumatori sia diminuita col passare degli anni; il grafico che illustra l'eventuale stagionalità presenta una marcata irregolarità, il che ci porta a dedurre l'assenza di una periodicità su base annuale del fenomeno d'interesse.

Figura 1.1: Percentuale fumatori complessiva; scomposizione in trend e stagionalità



1.2.1.2 Stratificazione: classe d'età

Suddividendo le persone intervistate in tre classi d'età (18-34, 35-49, 50-69), possiamo valutare l'andamento della percentuale di fumatori all'interno di ciascuno strato e confrontare tra loro e con la marginale le serie storiche ottenute. Prima di fare ciò, precisiamo che ogni mese, in media, gli intervistati appartenenti alla fascia 18-34 anni sono il 28.33%, la fascia 35-49 è rappresentata dal 34.55% degli individui, la fascia 50-69 dal 37.12%; la tabella 1.2 illustra più dettagliatamente le

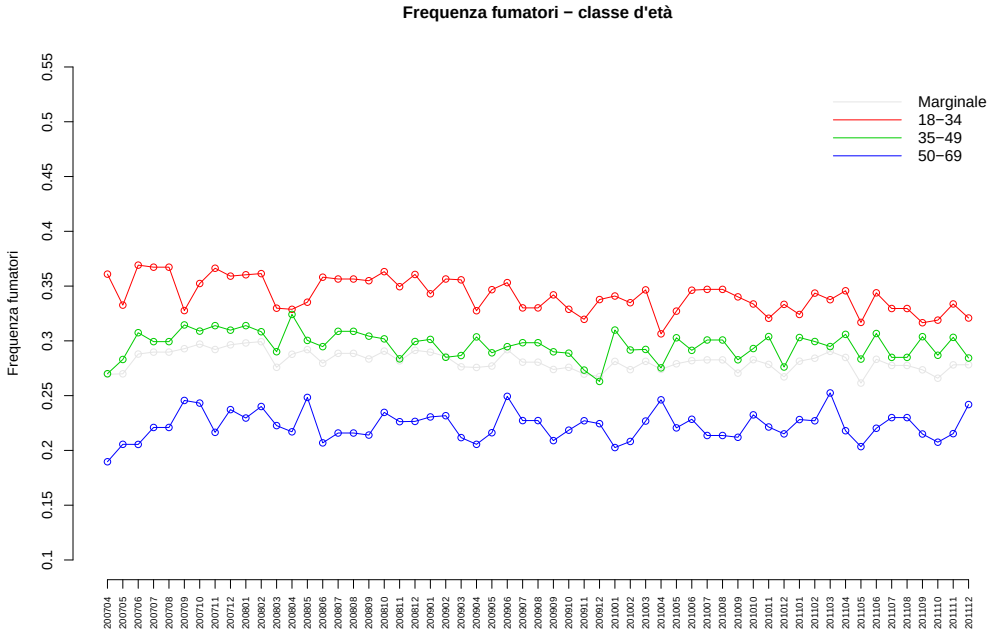
caratteristiche di ciascuno strato. Dal grafico 1.2 si può notare come in generale

Tabella 1.2: *Summary* degli strati ottenuti tramite il criterio della fascia d'età

	Dati mancanti	18-34	35-49	50-69
Min.	0	0.2592	0.3297	0.3437
1st Qu.	0	0.2739	10.3398	0.3616
Median	0	0.2833	0.3453	0.3698
Mean	0	0.2833	0.3455	0.3712
3rd Qu.	0	0.2908	0.3505	0.3791
Max.	0	0.3060	0.3642	0.4072

la percentuale di fumatori sia più alta tra i giovani, rispetto a coloro che appartengono alla fascia 35-49 e soprattutto rispetto agli individui tra i 50 e i 69 anni, la cui percentuale di fumatori è al di sotto della media globale.

Figura 1.2: Percentuale fumatori - stratificazione: classe d'età



Come prima, analizziamo ora le tre diverse serie (una per ogni strato) singolarmente per verificare la presenza o meno di una tendenza di fondo e/o di una componente stagionale. I grafici 1.3, 1.4, 1.5 mettono in evidenza la presenza di un trend complessivamente decrescente per quanto riguarda la percentuale

di fumatori registrata all'interno delle fasce 18-34 e 35-49, così come sembra improbabile la presenza di stagionalità all'interno di ciascuno di entrambi gli strati. Per quanto riguarda la terza classe d'età, invece, non constatiamo una tendenza di fondo regolare, ma possiamo notare una qualche regolarità nel grafico che mette in evidenza la stagionalità: in particolar modo, spiccano l'ascesa dai mesi invernali ai primaverili e il declino da questi ultimi ai mesi estivi.

Figura 1.3: Percentuale fumatori - classe d'età 18-34. Scomposizione in trend e stagionalità

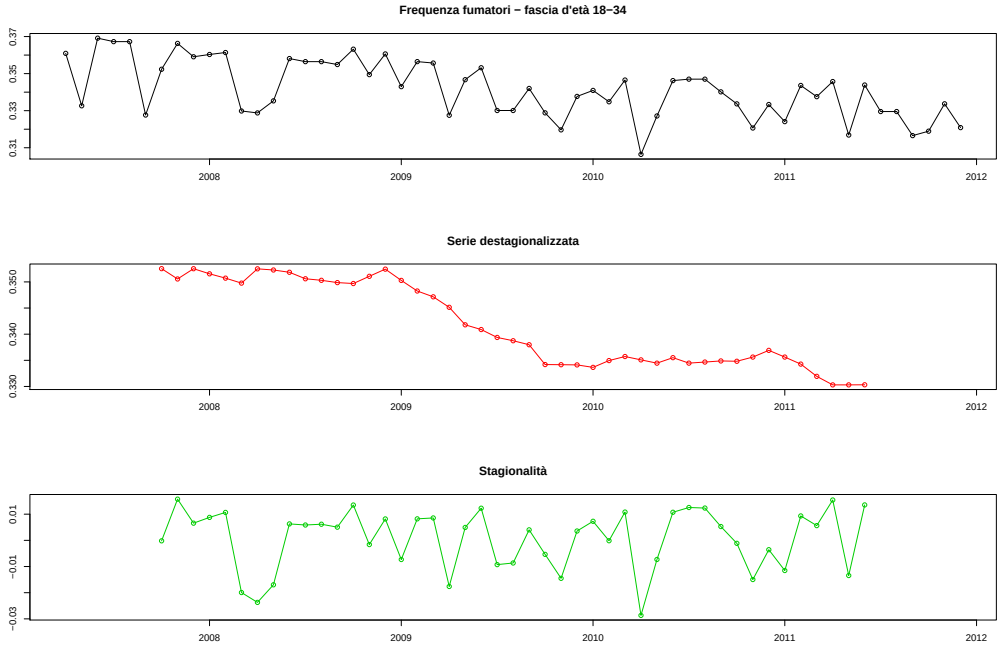


Figura 1.4: Percentuale fumatori - classe d'età 35-49. Scomposizione in trend e stagionalità

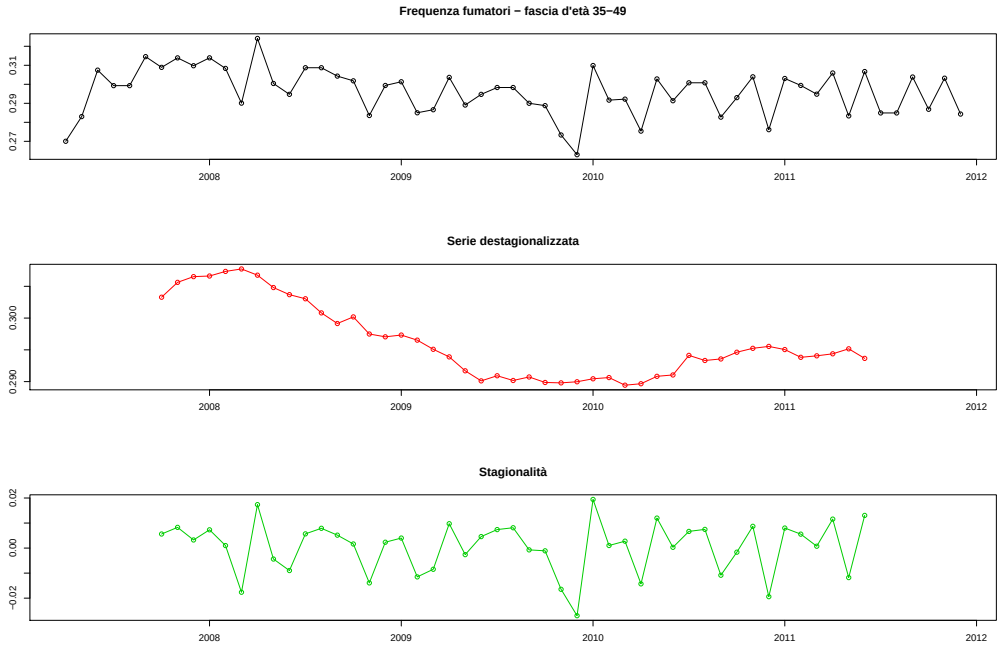
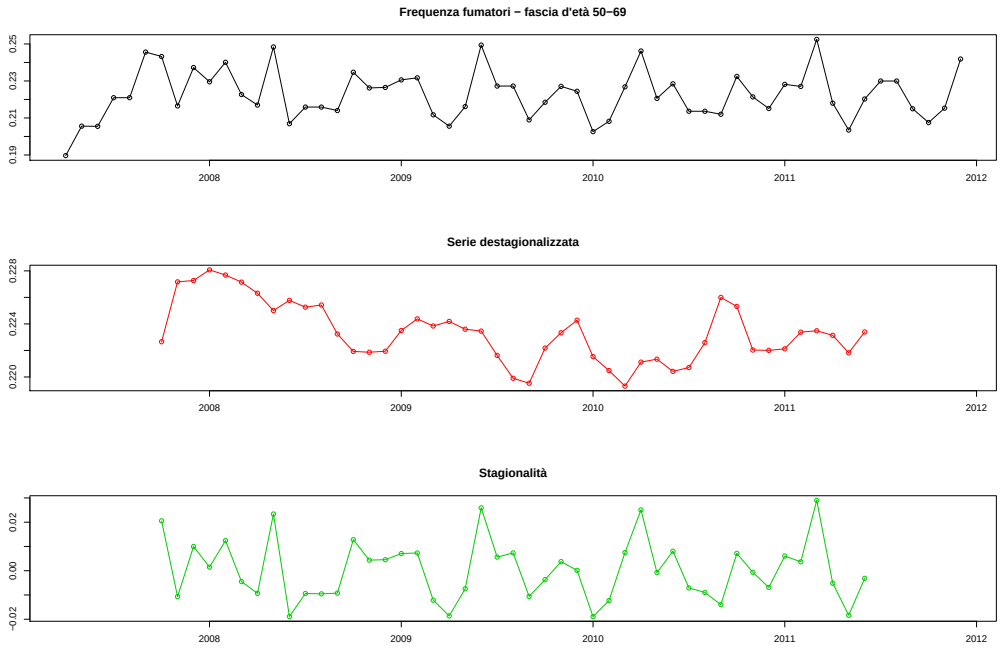


Figura 1.5: Percentuale fumatori - classe d'età 50-69. Scomposizione in trend e stagionalità



1.2.1.3 Stratificazione: sesso

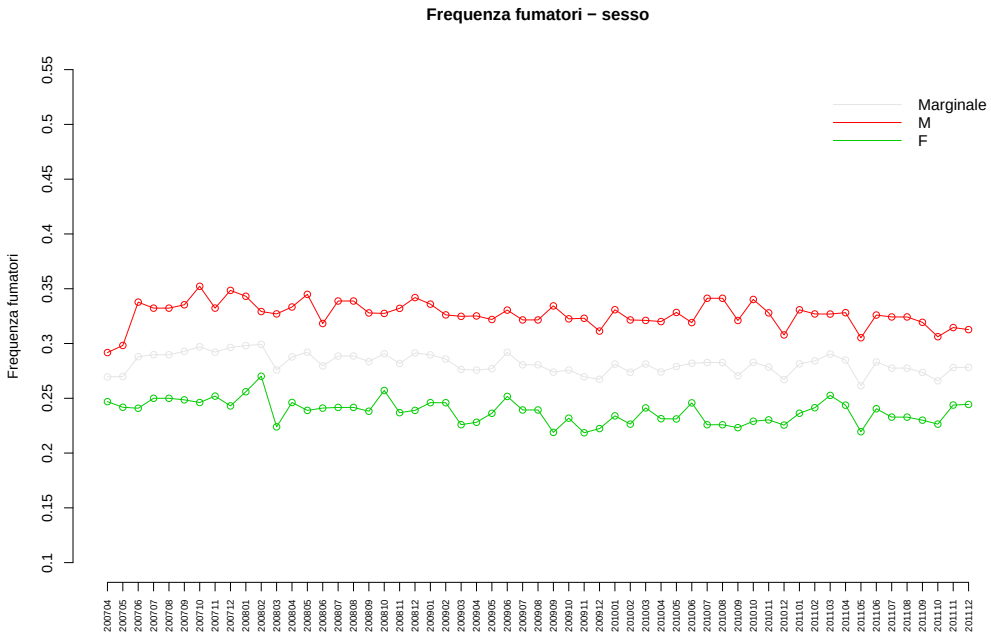
La tabella 1.3 illustra le caratteristiche dei campioni mensilmente rilevate, in base al sesso: mediamente, la percentuale di donne intervistate ogni mese supera, anche se non di molto, quella degli uomini. Il grafico 1.6 mostra come la propen-

Tabella 1.3: *Summary* degli strati ottenuti tramite il criterio del sesso

	Dati mancanti	M	F
Min.	0	0.4712	0.4889
1st Qu.	0	0.4840	0.4987
Median	0	0.4919	0.5081
Mean	0	0.4925	0.5075
3rd Qu.	0	0.5013	0.5160
Max.	0	0.5111	0.5288

sione al fumo sia più marcata nella fascia maschile della popolazione; possiamo inoltre notare una variabilità relativamente contenuta di entrambe le serie.

Figura 1.6: Percentuale fumatori - stratificazione: sesso



La serie riguardante lo strato maschile denota una tendenza di fondo complessivamente decrescente, fatta eccezione per un periodo di lieve ascesa tra il

2010 ed il 2011, e l'assenza di una componente stagionale regolare (1.7). Per quanto concerne l'andamento della serie riguardante le donne, riscontriamo un trend decrescente fino ai primi mesi del 2010 e successivamente crescente; anche in questo caso, non notiamo una stagionalità del fenomeno (1.8).

Figura 1.7: Percentuale fumatori - sesso maschile. Scomposizione in trend e stagionalità

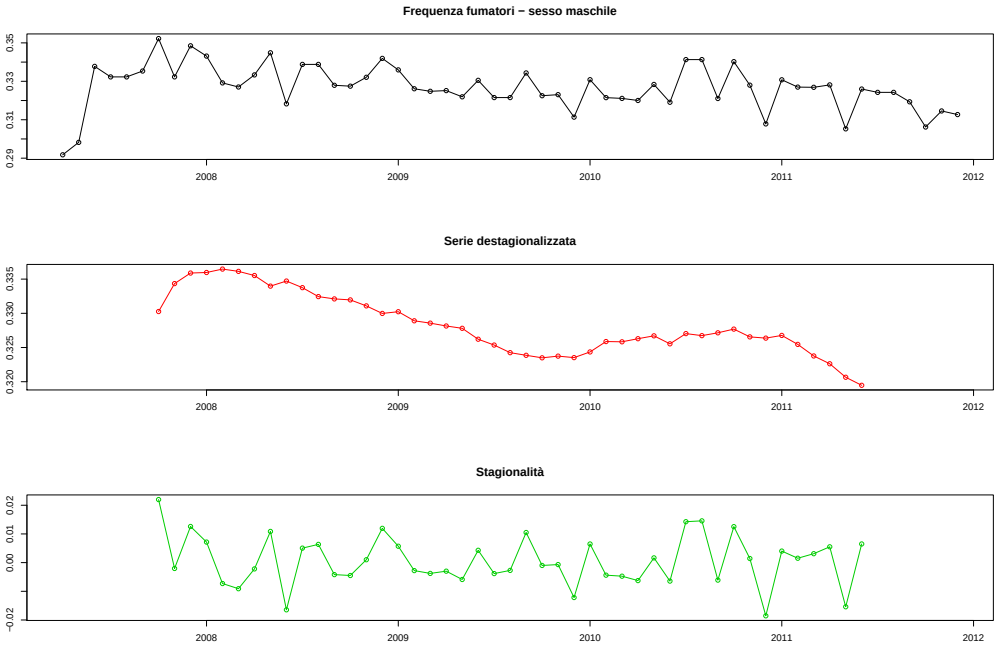
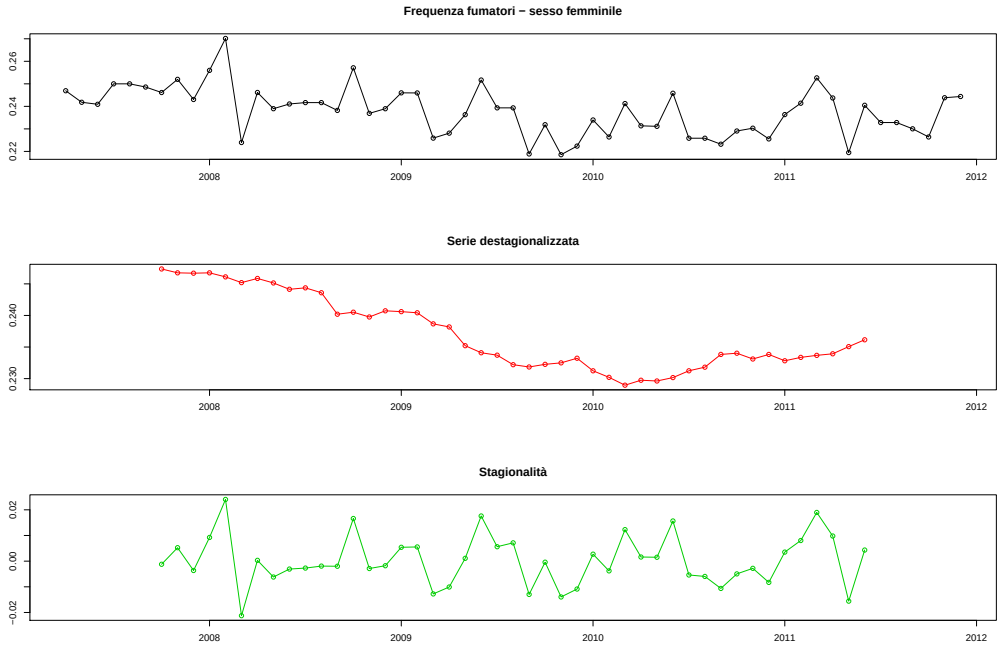


Figura 1.8: Percentuale fumatori - sesso femminile. Scomposizione in trend e stagionalità



1.2.1.4 Stratificazione: istruzione

Partizionando la popolazione in base al livello di istruzione, ogni mese in media il campione intervistato è composto per quasi il 12% da individui aventi nessun titolo di studio oppure la licenza elementare, da circa il 31% di diplomati alla scuola media inferiore, da più del 43% di diplomati alla scuola media superiore e da quasi il 13% di persone laureate (tabella 1.4). Dal grafico 1.9 notiamo

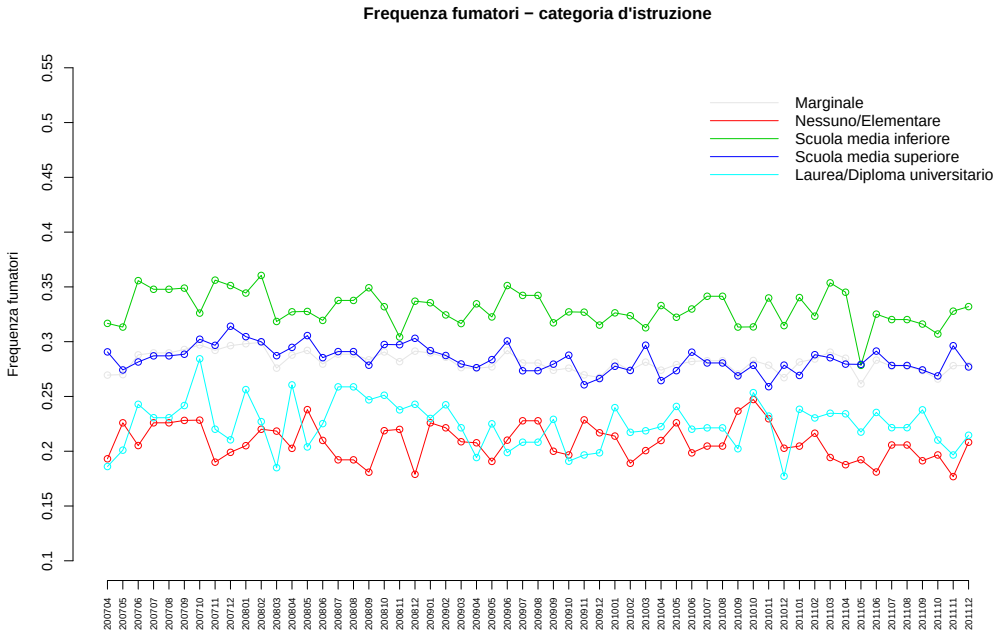
Tabella 1.4: *Summary* degli strati ottenuti tramite il criterio del livello di istruzione

	Dati mancanti	Nessuno/Elem.	Media inf.	Media sup.	Laurea
Min.	0.0003	0.0899	0.2847	0.3919	0.1150
1st Qu.	0.0009	0.1055	0.3086	0.4268	0.1246
Median	0.0013	0.1201	0.3130	0.4372	0.1283
Mean	0.0015	0.1191	0.3125	0.4372	0.1297
3rd Qu.	0.0018	0.1295	0.3165	0.4465	0.1346
Max.	0.0034	0.1522	0.3283	0.4664	0.1587

che la serie storica che si riferisce alla fascia di popolazione avente il diploma

di scuola media inferiore giace costantemente al di sopra della serie marginale; all'opposto, le serie che ricalcano l'andamento della percentuale di fumatori tra la fascia 'nessun titolo/licenza elementare' e 'laurea/diploma universitario' sono tracciate al di sotto del livello medio globale. La percentuale di fumatori tra coloro che possiedono il diploma di scuola media superiore è molto vicina al valore marginale.

Figura 1.9: Percentuale fumatori - stratificazione: istruzione



Anche questo criterio di stratificazione non evidenzia la presenza di una componente stagionale, come testimoniano i grafici 1.10, 1.11, 1.12, 1.13, che mettono in luce alcune tendenze di fondo per quanto riguarda solo alcune fasce.

Figura 1.10: Percentuale fumatori - livello di istruzione: nessuno/elementare. Scomposizione in trend e stagionalità

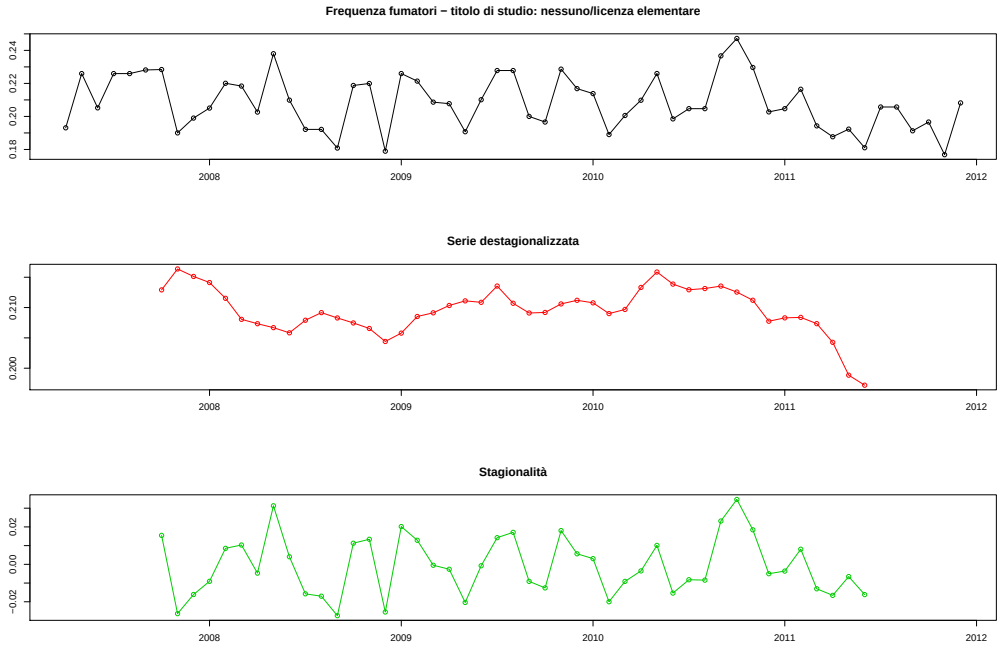


Figura 1.11: Percentuale fumatori - livello di istruzione: scuola media inferiore. Scomposizione in trend e stagionalità

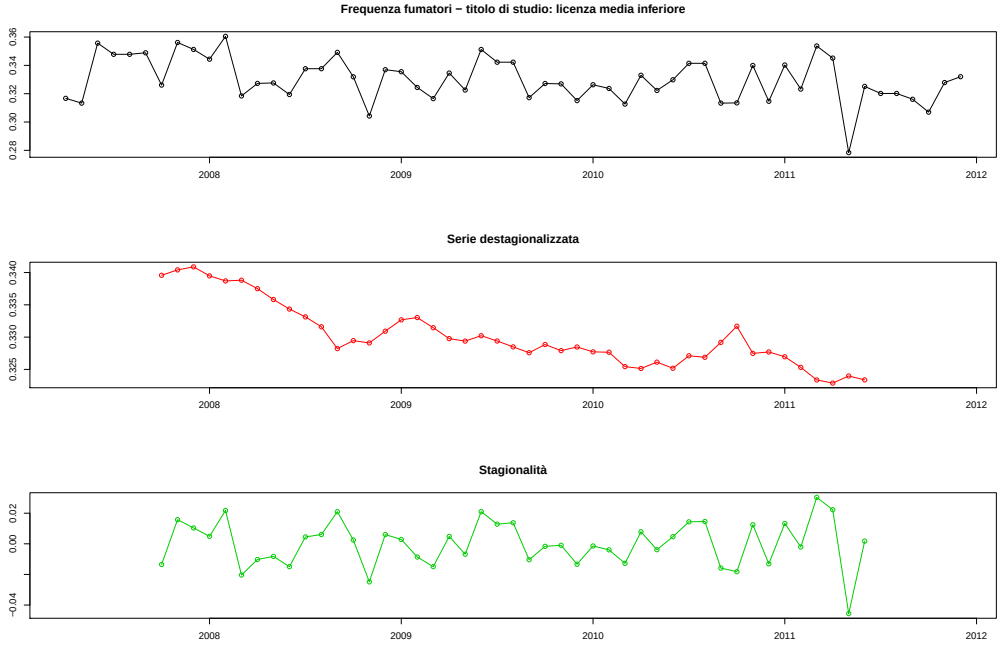


Figura 1.12: Percentuale fumatori - livello di istruzione: scuola media superiore. Scomposizione in trend e stagionalità

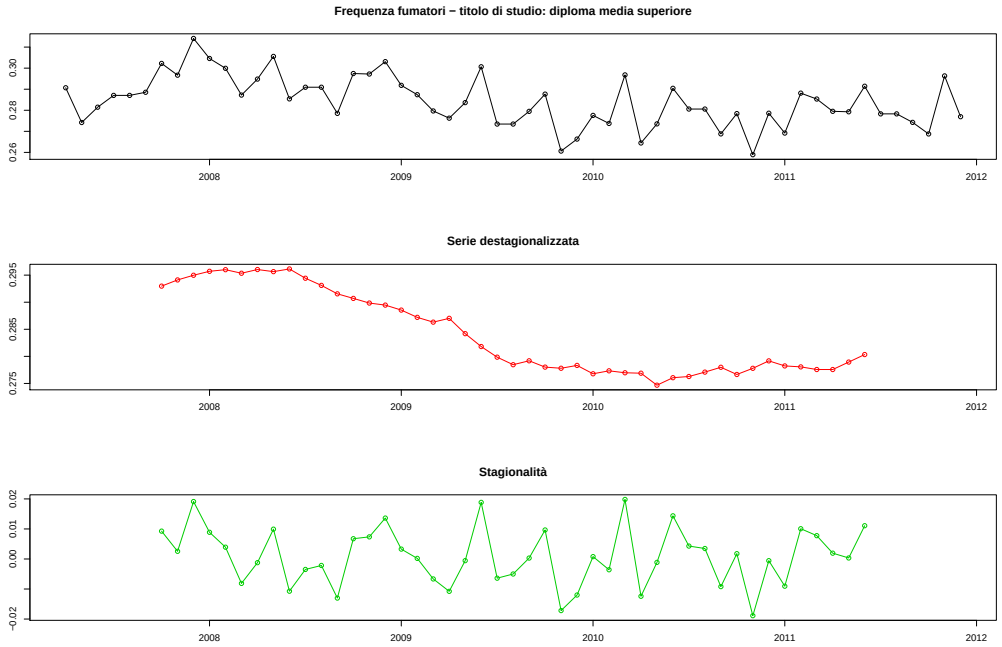
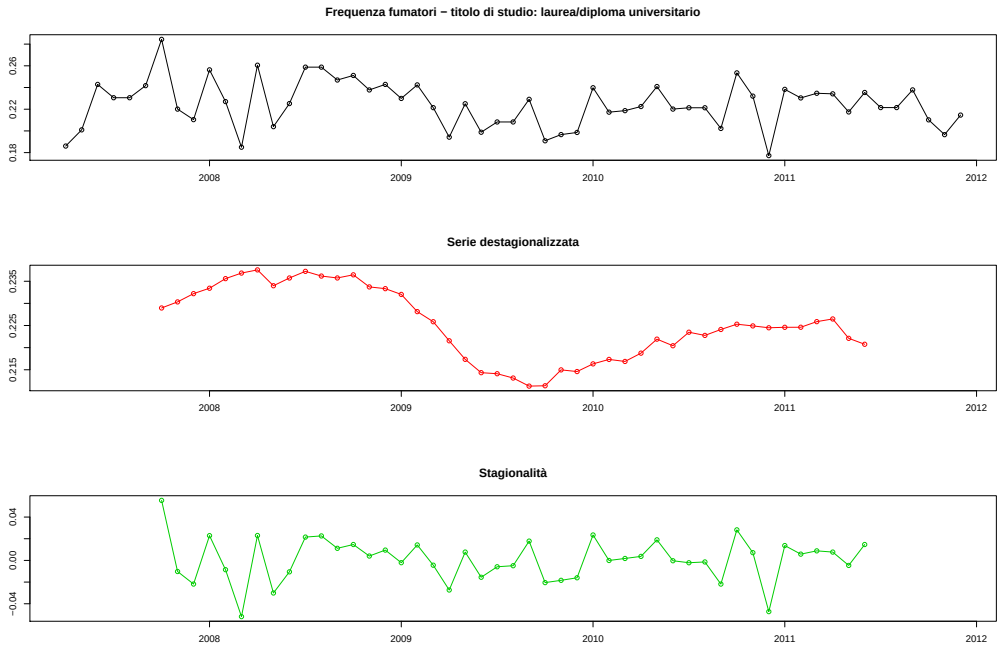


Figura 1.13: Percentuale fumatori - livello di istruzione: laurea/diploma universitario. Scomposizione in trend e stagionalità



1.2.1.5 Stratificazione: regione

Operiamo in questo caso una stratificazione territoriale degli intervistati, suddividendoli nelle 21 regioni (il Trentino Alto Adige viene scisso nelle province autonome di Bolzano e Trento) italiane. Di seguito sono riportati i summary di ogni regione con indicazioni sulla numerosità campionaria e sulla percentuale di fumatori (tabelle 1.5, 1.6, 1.7, 1.8, 1.9, 1.10, 1.11, 1.12, 1.13, 1.14, 1.15, 1.16, 1.17, 1.18, 1.19, 1.20, 1.21, 1.22, 1.23, 1.24, 1.25) ed i grafici delle serie storiche (figure 1.14, 1.15, 1.16, 1.17, 1.18, 1.19, 1.20), che presentano situazioni variegate: dati mancanti perchè in alcune regioni non sono state svolte le rilevazioni in tutti i mesi considerati, maggiore variabilità delle serie riguardanti le regioni aventi campioni meno numerosi, tendenze di fondo di diverso tipo e stagionalità pressochè scarsa, almeno ad una prima occhiata.

Tabella 1.5: *Summary* della regione Piemonte

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	1	0	0	0
1st Qu.	333	0	0.7039	0.2624
Median	368	0	0.7201	0.2792
Mean	403.1	0.0182	0.7061	0.2757
3rd Qu.	490	0	0.7349	0.2927
Max.	611	1	0.7631	0.3303
NA's	1	1	1	1

Tabella 1.6: *Summary* della regione Valle d'Aosta

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	25	0	0.5517	0.07143
1st Qu.	27.5	0	0.6774	0.21825
Median	29	0	0.7333	0.26667
Mean	28.76	0.002488	0.7350	0.26249
3rd Qu.	30	0	0.7778	0.32258
Max.	31	0.035714	0.9286	0.44828
NA's	2	2	2	2

Tabella 1.7: *Summary* della regione Lombardia

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	25	0	0.6027	0.0800
1st Qu.	121.2	0	0.7011	0.2540
Median	128	0	0.7225	0.2775
Mean	123.1	0.000621	0.7227	0.2767
3rd Qu.	131.8	0	0.7399	0.2989
Max.	147	0.015385	0.9200	0.3973
NA's	7	7	7	7

Tabella 1.8: *Summary* della provincia autonoma di Bolzano

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	22	0	0.5200	0.1200
1st Qu.	24	0	0.6667	0.2290
Median	25	0	0.7273	0.2727
Mean	26.47	0.002126	0.7247	0.2731
3rd Qu.	26	0	0.7600	0.3333
Max.	36	0.04	0.8800	0.4800
NA's	2	2	2	2

Tabella 1.9: *Summary* della provincia autonoma di Trento

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	42	0	0.5686	0.1304
1st Qu.	58	0	0.7108	0.200
Median	67	0	0.7619	0.2381
Mean	66.39	0.0005605	0.7564	0.243
3rd Qu.	75	0	0.800	0.2892
Max.	93	0.0172414	0.8696	0.4314

Tabella 1.10: *Summary* della regione Veneto

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	209	0	0.7171	0.1992
1st Qu.	443	0	0.7446	0.2265
Median	475	0	0.7598	0.2402
Mean	459.8	0.001197	0.7593	0.2395
3rd Qu.	502	0.002016	0.7727	0.2534
Max.	579	0.008772	0.8008	0.2809

Tabella 1.11: *Summary* della regione Friuli Venezia Giulia

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	104	0	0.6442	0.2118
1st Qu.	147	0	0.7040	0.2493
Median	163	0	0.7241	0.2759
Mean	159.3	0.0005572	0.7253	0.2742
3rd Qu.	169.5	0	0.7507	0.2925
Max.	187	0.0067568	0.7882	0.3558
NA's	2	2	2	2

Tabella 1.12: *Summary* della regione Liguria

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	114	0	0.600	0.2231
1st Qu.	125	0	0.6917	0.2578
Median	127	0	0.7165	0.2835
Mean	127.7	0.0005653	0.7157	0.2837
3rd Qu.	130	0	0.7422	0.3077
Max.	147	0.0087719	0.7769	0.400

Tabella 1.13: *Summary* della regione Emilia Romagna

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	116	0	0.6189	0.2354
1st Qu.	300	0	0.6885	0.2778
Median	334	0	0.7014	0.2977
Mean	325.2	0.000287	0.7007	0.2990
3rd Qu.	346	0	0.7222	0.3115
Max.	413	0.007937	0.7566	0.3811

Tabella 1.14: *Summary* della regione Toscana

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	2	0	0.6667	0
1st Qu.	255	0	0.6927	0.2747
Median	291	0	0.7148	0.2830
Mean	321.6	0.001950	0.7170	0.2811
3rd Qu.	332	0.001116	0.7242	0.3021
Max.	981	0.028490	1	0.3275
NA's	1	1	1	1

Tabella 1.15: *Summary* della regione Umbria

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	35	0	0.5971	0.1944
1st Qu.	113	0	0.6542	0.2734
Median	127	0	0.6822	0.3178
Mean	121.9	0	0.6878	0.3122
3rd Qu.	129	0	0.7266	0.3458
Max.	158	0	0.8056	0.4029

Tabella 1.16: *Summary* della regione Marche

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	25	0	0.6417	0.1607
1st Qu.	93	0	0.6939	0.2481
Median	113	0	0.7174	0.2812
Mean	110	0.001136	0.7231	0.2758
3rd Qu.	131	0	0.7519	0.3045
Max.	187	0.020833	0.8393	0.3529
NA's	2	2	2	2

Tabella 1.17: *Summary* della regione Lazio

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	148	0	0.5847	0.2603
1st Qu.	259.5	0	0.6554	0.3013
Median	277.5	0	0.6742	0.3241
Mean	272.9	0.001383	0.6786	0.3200
3rd Qu.	298	0	0.6982	0.3394
Max.	337	0.020080	0.7363	0.4153
NA's	3	3	3	3

Tabella 1.18: *Summary* della regione Abruzzo

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	43	0	0.5962	0.2162
1st Qu.	76	0	0.6522	0.2787
Median	96	0	0.6892	0.3108
Mean	98.61	0.0002114	0.6882	0.3116
3rd Qu.	122	0	0.7213	0.3478
Max.	148	0.0120482	0.7838	0.4038

Tabella 1.19: *Summary* della regione Molise

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	23	0	0.5200	0.1724
1st Qu.	25	0	0.6800	0.2414
Median	27	0	0.7200	0.2703
Mean	30.73	0.002537	0.7102	0.2872
3rd Qu.	37	0	0.7500	0.3200
Max.	47	0.040000	0.8276	0.4800
NA's	12	12	12	12

Tabella 1.20: *Summary* della regione Campania

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	140	0	0.6219	0.2273
1st Qu.	193	0	0.6758	0.2833
Median	258	0	0.6960	0.3040
Mean	252.2	0.00005907	0.6982	0.3018
3rd Qu.	305	0	0.7167	0.3242
Max.	409	0.003367	0.7727	0.3781

Tabella 1.21: *Summary* della regione Puglia

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	99	0	0.6548	0.18424
1st Qu.	174	0	0.6816	0.2591
Median	187	0	0.7090	0.2902
Mean	182.9	0.00202	0.7111	0.2869
3rd Qu.	197.2	0	0.7386	0.3153
Max.	262	0.02139	0.8158	0.3452
NA's	5	5	5	5

Tabella 1.22: *Summary* della regione Basilicata

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	22	0	0.6250	0.1224
1st Qu.	91	0	0.7183	0.2204
Median	119.5	0	0.7458	0.2498
Mean	113.3	0.004364	0.7517	0.2439
3rd Qu.	145.8	0	0.7764	0.2748
Max.	192	0.058824	0.8776	0.3750
NA's	11	11	11	11

Tabella 1.23: *Summary* della regione Calabria

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	28	0	0.5909	0.1216
1st Qu.	57	0	0.7073	0.2295
Median	67	0	0.7419	0.2535
Mean	66.06	0.001855	0.7383	0.2599
3rd Qu.	82	0	0.7705	0.2927
Max.	99	0.026316	0.8784	0.4091
NA's	24	24	24	24

Tabella 1.24: *Summary* della regione Sicilia

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	75	0	0.6117	0.2203
1st Qu.	108	0	0.6769	0.2650
Median	124	0	0.7107	0.2893
Mean	123.6	0.0005245	0.7041	0.2954
3rd Qu.	139	0	0.7350	0.3231
Max.	195	0.0097087	0.7797	0.3871

Tabella 1.25: *Summary* della regione Sardegna

	Numerosità	Dati mancanti	Non fumatori	Fumatori
Min.	17	0	0.6216	0.07407
1st Qu.	44	0	0.7071	0.2235
Median	59	0	0.7482	0.2517
Mean	63.31	0.000687	0.7450	0.2543
3rd Qu.	78.25	0	0.7765	0.2927
Max.	140	0.035714	0.9259	0.3783
NA's	5	5	5	5

Figura 1.14: Percentuale fumatori - Piemonte, Valle d'Aosta, Lombardia. Scomposizione in trend e stagionalità

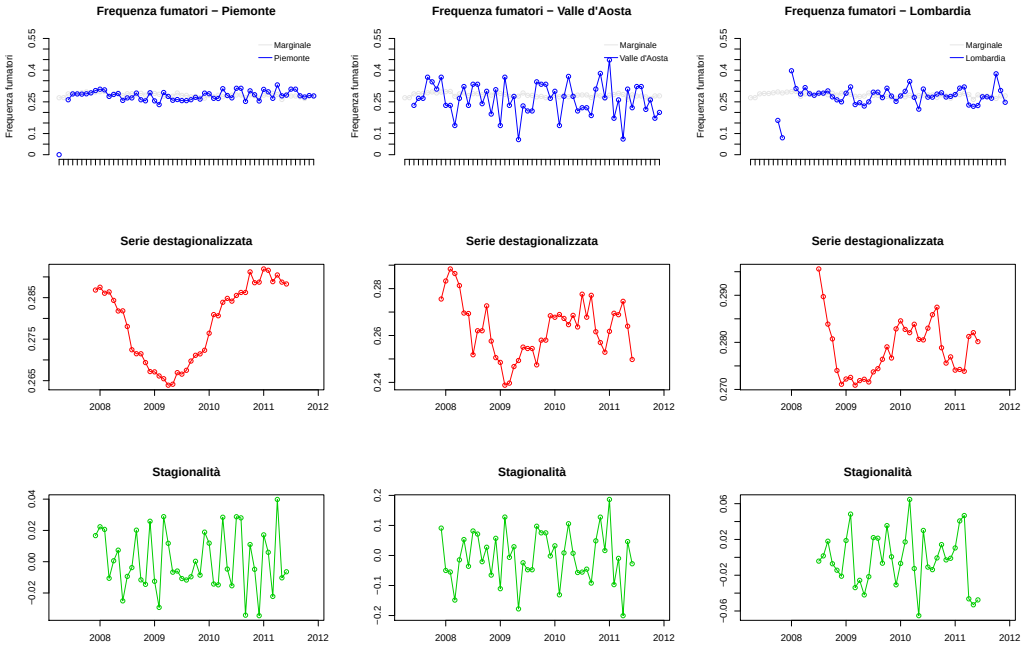


Figura 1.15: Percentuale fumatori - Bolzano, Trento, Veneto. Scomposizione in trend e stagionalità

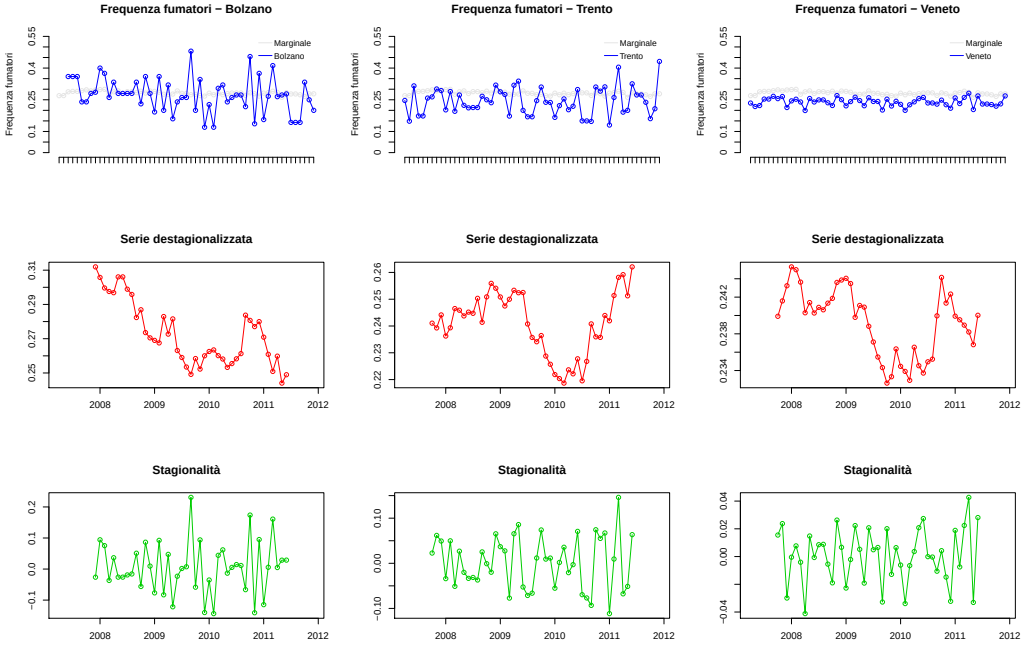


Figura 1.16: Percentuale fumatori - Friuli Venezia Giulia, Liguria, Emilia Romagna. Scomposizione in trend e stagionalità

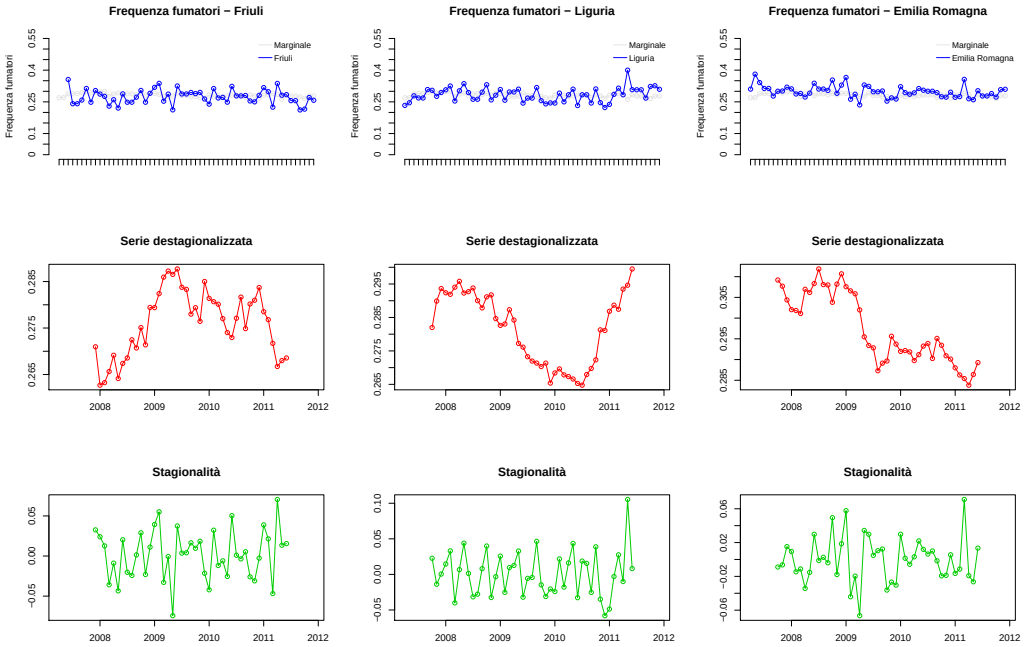


Figura 1.17: Percentuale fumatori - Toscana, Umbria, Marche. Scomposizione in trend e stagionalità

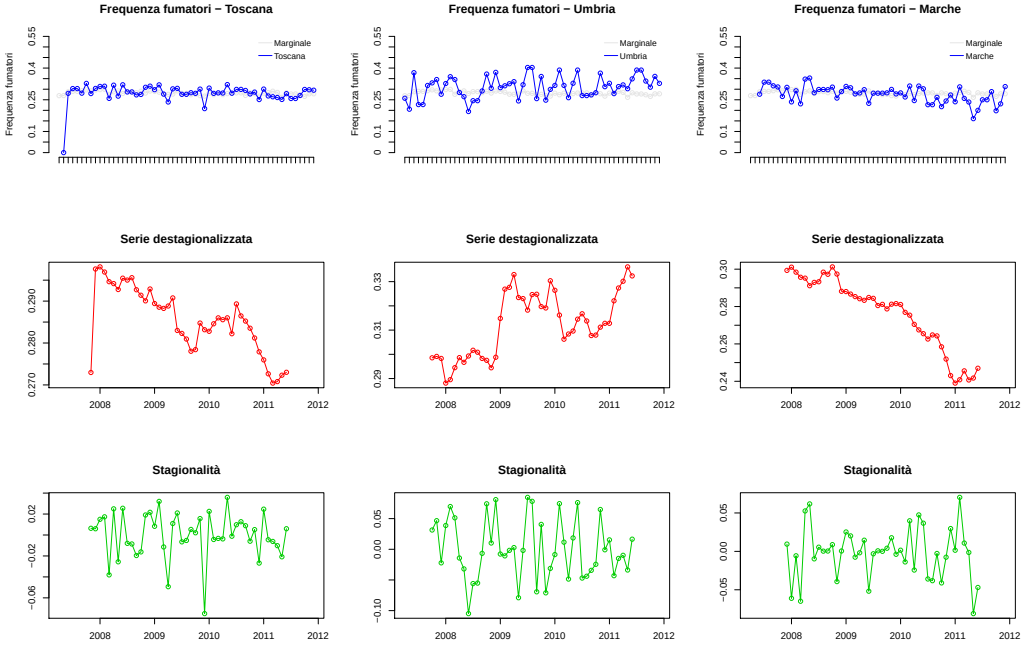


Figura 1.18: Percentuale fumatori - Lazio, Abruzzo, Molise. Scomposizione in trend e stagionalità

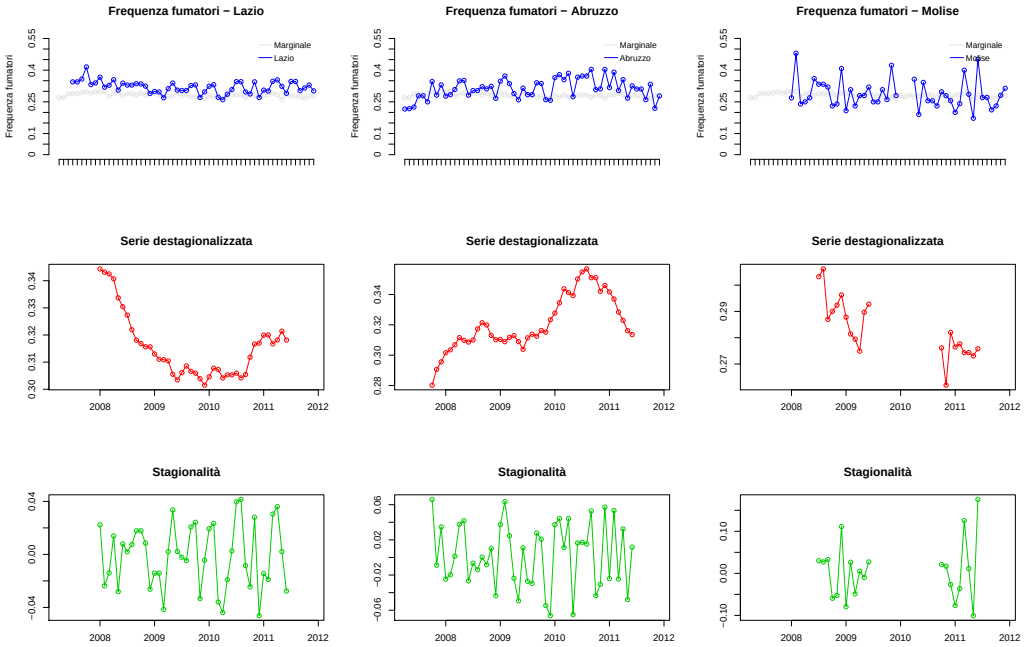


Figura 1.19: Percentuale fumatori - Campania, Puglia, Basilicata. Scomposizione in trend e stagionalità

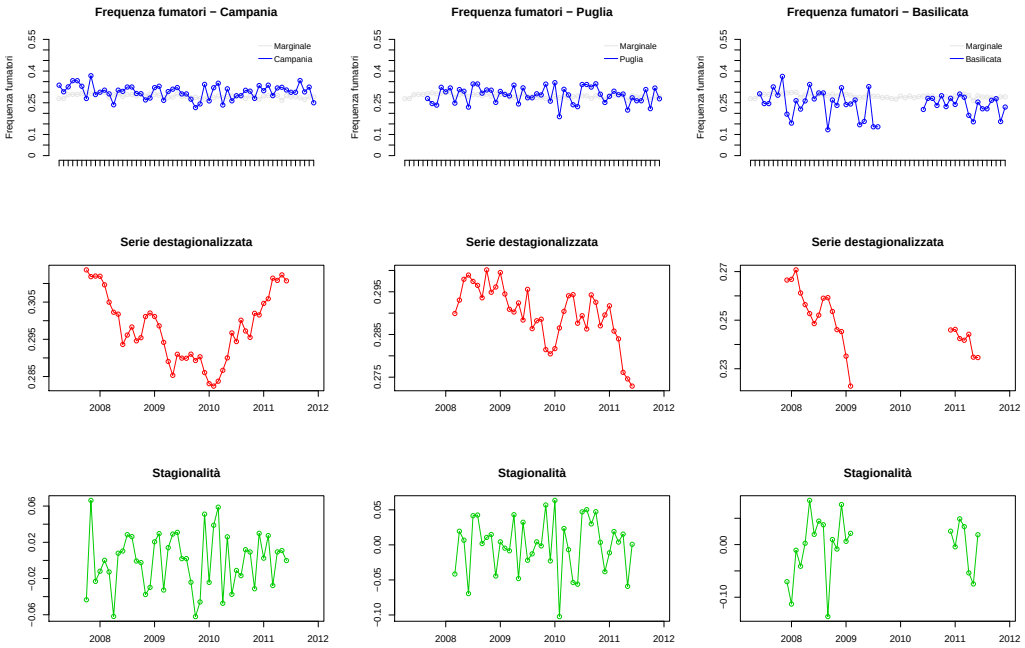
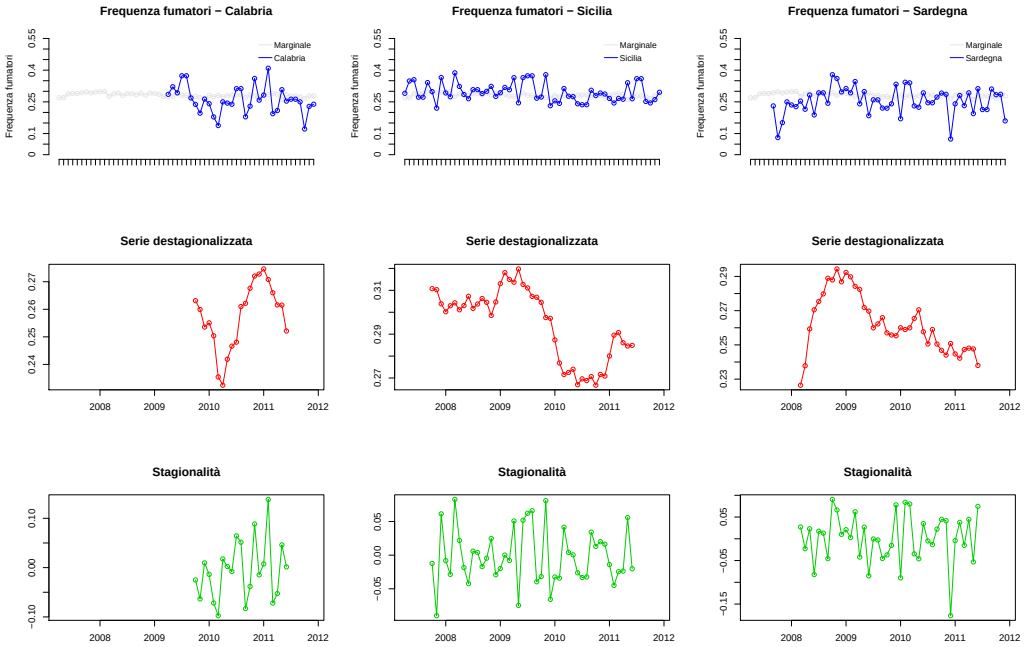


Figura 1.20: Percentuale fumatori - Calabria, Sicilia, Sardegna. Scomposizione in trend e stagionalità



Poichè per il mese di aprile 2007 mancano le rilevazioni di ben 12 regioni ed altrettante per il mese di maggio 2007, decidiamo di non prendere in considerazione questi due periodi nelle successive analisi. Dopo questa esclusione, 7 regioni presentano ancora dei mesi senza osservazioni a disposizione.
Nel capitolo successivo ci occuperemo di questo aspetto, offriremo un'analisi più dettagliata delle serie storiche ed inizieremo ad occuparci della modellazione.

Capitolo 2

Modellazione: GLM per serie storiche

2.1 Modelli di regressione lineari generalizzati

La classe dei modelli lineari generalizzati permette di superare alcune delle limitazioni dei modelli lineari classici, come spiegato nel seguito [Azzalini 2001], [Pezzato 2010]:

1. supponendo che la variabile Y risulti dalla somma di due termini del tipo $y = r(x_1, \ldots, x_n) + \varepsilon$, può succedere che la componente sistematica $r(\cdot)$ sia nota e decisamente non lineare nei parametri;

2. anche quando $r(\cdot)$ non è di forma nota, talvolta sappiamo abbastanza sulla natura del fenomeno da escludere a priori una relazione lineare;

3. alcune situazioni richiedono un campo di variazione limitato per il valore atteso $E(Y)$;

4. la varianza del termine d'errore, e quindi anche della variabile risposta, può non essere costante, ma essere, per esempio, funzione di $E(Y)$;

5. la distribuzione della variabile risposta può non essere Normale e potrebbe non essere applicabile una trasformazione non lineare dei dati per ricondurli alla normalità.

2.1.1 Formulazione dei modelli lineari generalizzati

Descriveremo la classe dei modelli lineari generalizzati partendo da una categoria di distribuzioni facente parte della famiglia esponenziale. Per una variabile casuale Y scriveremo

$$Y \sim \mathcal{EF}(b(\theta), \phi)$$

per indicare che Y ha funzione di densità del tipo

$$f(y; \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\},$$

dove θ è un parametro scalare ignoto, ϕ è un parametro di dispersione che possiamo considerare fissato, $b(\cdot)$ e $c(\cdot)$ sono funzioni note la cui scelta individua una particolare distribuzione di probabilità [Ventura 2008].

Diverse notevoli distribuzioni appartengono alla famiglia esponenziale, quali le distribuzioni Normale, Poisson, Binomiale. Per ottenerle è necessario scegliere opportunamente le funzioni $b(\cdot)$ e $c(\cdot)$.

Dal calcolo dei momenti otteniamo le formulazioni del valore atteso di Y, $E(Y) = b'(\theta) = \mu$, e della sua varianza, $var(Y) = b''(\theta) = \phi V(\mu)$.

Un modello appartenente alla classe dei GLM (acronimo dell'inglese *Generalized Linear Models*) è caratterizzato dalle seguenti componenti:

- la componente casuale: $Y_i \sim \mathcal{EF}(b(\theta_i), \phi)$, indipendenti, con $E(Y_i) = \mu_i = b'(\theta_i), i = 1, 2, \ldots, n$;
- la componente sistematica: $\eta_i = x_i^T \beta$, dove x_i è un vettore di costanti (variabile esplicativa) e β è un vettore di parametri;
- la funzione legame $g(\cdot)$, formulata in modo tale che $g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i), i = 1, 2, \ldots, n$.

2.1.2 Stime di massima verosimiglianza

La stima dei parametri nei GLM viene effettuata tramite la procedura di massima verosimiglianza. Sfruttando l'ipotesi di indipendenza tra le Y_i, la log-verosimiglianza risulta

$$l(\beta) = \sum_{i=1}^{n} \ln f(y_i; \theta_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} = \sum_{i=1}^{n} l_i(\beta)$$

con $g(\mu_i) = g(b'(\theta_i)) = \eta_i = x_i^T\beta$. Per ottenere la stima di β è necessario risolvere, attraverso metodi iterativi, le equazioni di verosimiglianza

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0, \qquad j = 1, 2, \ldots, p$$

che assumono la seguente formulazione:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi V(\mu_i) g'(\mu_i)} x_{ij} = 0, \qquad j = 1, 2, \ldots, p.$$

La soluzione è la stessa per qualsiasi valore di ϕ, sia esso noto oppure no; in quest'ultimo caso si potrebbe ricorrere alla massima verosimiglianza per stimare il parametro di dispersione, ma è prassi utilizzare un altro stimatore, più stabile e robusto:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

dove $\hat{\mu}_i = g^{-1}(x_i^T\hat{\beta})$ sono i valori di μ_i stimati, messi a disposizione dall'algoritmo iterativo per β.

Per n sufficientemente grande, $\hat{\beta}$ si distribuisce approssimativamente come una Normale multivariata con media β e matrice di varianze e covarianze $\hat{\phi}(X^T\hat{W}X)^{-1}$, dove $\hat{W}$ proviene dal calcolo della stima di massima verosimiglianza di β.

Di conseguenza, per un parametro β_j, possiamo costruire l'intervallo di confidenza approssimato di livello $1-\alpha$ in questo modo:

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\phi[(X^T\hat{W}X)^{-1}]_{i,j}}.$$

Se ϕ non è noto, viene stimato ed il quantile $z_{1-\alpha/2}$ è sostituito da $t_{n-p,1-\alpha/2}$.

2.1.3 Valutazione della bontà del modello

Valutare l'adeguatezza di un modello significa, in poche parole, stabilire quanto discrepanti siano l'insieme di valori stimati ($\hat{\mu}$) e l'insieme dei dati (y). Per poter quantificare questa discrepanza ci affidiamo alla devianza, una misura basata sulla log-verosimiglianza.

Supponiamo di voler confrontare M_C e M_R, due modelli lineari generalizzati annidati tali per cui $M_R \subset M_C$; M_C contiene p parametri e M_R ne contiene p_0, con $p_0 > p$. Costruiamo una partizione di β: $\beta = (\beta_{MR}, \beta_{MRC})$, con $\beta_{MR} = (\beta_1, \ldots, \beta_{p_0})$ e $\beta_{MRC} = (\beta_{p_0+1}, \ldots, \beta_p)$ e per effettuare la verifica d'ipotesi

$$\begin{cases} H_0 : \beta_{MRC} = 0 \\ H_1 : \beta_{MRC} \neq 0 \end{cases}$$

ci affidiamo al rapporto di verosimiglianza

$$W = 2\{l(\hat{\beta}) - l(\hat{\beta_{MR}})\}.$$

A questo punto definiamo la funzione di devianza scalata del modello:

$$D^*(y; \hat{\theta}) = \frac{2\phi\{l(\tilde{\theta}) - l(\hat{\theta})\}}{\phi},$$

dove $l(\tilde{\theta})$ è la funzione di log-verosimiglianza del modello saturo, ossia del modello basato sulla stessa distribuzione del modello corrente preso in considerazione, avente la sua stessa funzione legame e contenente un numero di parametri pari al numero di osservazioni, n. Se il modello corrente si adatta bene ai dati, $l(\hat{\theta})$ si discosta poco da $l(\tilde{\theta})$ e la devianza scalata risulta piccola.

Nel caso di due modelli annidati, dunque, il rapporto di verosimiglianza W può essere riscritto come

$$W = D^*(Y, \hat{\theta_{MR}}) - D^*(Y, \hat{\theta}),$$

quantità che sotto l'ipotesi nulla e con ϕ noto, si distribuisce approssimativamente come una $\chi^2_{p-p_0}$; si rifiuta H_0 se la statistica W assume valori elevati, che superano il quantile di livello $1 - \alpha$ della distribuzione, e quindi se il *p-value* è

ragionevolmente piccolo. Se ϕ non è noto, lo si sostituisce con la sua stima e, approssimativamente,

$$\frac{W}{p-p_0} = \frac{D^*(Y,\hat{\theta_{MR}}) - D^*(Y,\hat{\theta})}{p-p_0}$$

si distribuisce come una $F_{p-p_0,n-p}$.

Un ulteriore strumento per verificare la correttezza della specificazione del modello è l'analisi dei residui. Definiamo il residuo di Pearson:

$$r_{P_i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}V(\hat{\mu}_i)}},$$

il quale, se il modello è valido, dovrebbe approssimativamente seguire una distribuzione $N(0,1)$.

2.1.4 Distribuzione Binomiale

Se il nostro obiettivo è descrivere i risultati (successi/insuccessi) di prove indipendenti, ci affidiamo alla distribuzione Binomiale, facente parte della famiglia esponenziale.

Sia $Z \sim Bin(m,\pi)$; se consideriamo $Y = Z/m$ e definiamo

$$\theta = \ln\frac{\pi}{1-\pi}, \qquad \phi = \frac{1}{m},$$

otteniamo

$$b(\theta) = \ln(1+e^\theta), \qquad c(y,\phi) = \ln\binom{1/\phi}{y/\phi}.$$

Dunque $Y \sim \mathcal{EF}(\ln(1+e^\theta), 1/m)$.

Il valore atteso di Y è pari a

$$E(Y) = \mu = b'(\theta) = \frac{e^\theta}{1+e^\theta} = \pi$$

e la varianza è

$$var(Y) = \frac{1}{m}\pi(1-\pi),$$

con $V(\mu) = \mu(1-\mu)$.

Per la specificazione completa di un GLM abbiamo bisogno della funzione legame che metta in relazione il valore atteso delle Y_i e le variabili esplicative ($\eta_i = x_i^T\beta$). Nel caso in questione vogliamo modellare la probabilità π_i e, poichè $E(Y_i) = \mu_i = \pi_i$, il parametro μ_i dev'essere contenuto nell'intervallo $[0, 1]$. Dobbiamo dunque scegliere una funzione legame che ci permetta di rispettare questo vincolo; una di esse è la funzione logistica:

$$g(\mu_i) = logit(\mu_i) = \ln \frac{\mu_i}{1-\mu_i} = \eta_i = x_i^T\beta;$$

invertendo la relazione, otteniamo l'espressione per la probabilità di successo:

$$\mu_i = \frac{e^{x_i^T\beta}}{1 + e^{x_I^T\beta}}.$$

2.2 Quasi-verosimiglianza

Nel caso in cui l'ipotesi che le osservazioni Y_i provengano dalla famiglia esponenziale non sia soddisfatta, possiamo pervenire a delle stime usando il metodo della quasi-verosimiglianza, specificato dalle seguenti assunzioni [Ventura 2008]:

1. $g(\mu_i) = g(E(Y_i)) = \eta_i, \qquad 1 = 1, \ldots, n;$
2. $var(Y_i) = \phi V(\mu_i), \qquad 1 = 1, \ldots, n;$
3. indipendenza delle osservazioni.

Le equazioni di quasi-verosimiglianza permettono di definire una funzione di quasi-verosimiglianza per β,

$$l_Q(\beta) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)}\, dt,$$

che non è una funzione di log-verosimiglianza ma ne possiede molte proprietà. La quasi-devianza scalata,

$$D_Q^*(y; \hat{\beta}) = \frac{-2\phi l_Q(\hat{\beta})}{\phi},$$

può essere utilizzata per confrontare i modelli annidati e segue distribuzioni analoghe a quelle descritte in precedenza per la devianza; anche la stima per il parametro di dispersione resta la stessa, così come la distribuzione dello stimatore $\hat{\beta}$.

2.2.1 Sovradispersione e sottodispersione

A volte, soprattutto trattando dati discreti, la varianza della variabile Y ossia delle osservazioni, è maggiore o minore di quella teorica, ossia dei valori stimati dal modello; ci troviamo dunque di fronte a problemi rispettivamente di sovradispersione o sottodispersione. In questo caso i dati ci suggeriscono un valore di ϕ maggiore o minore di 1, mentre le distribuzioni discrete facenti parte della famiglia esponenziale (Poisson e Binomiale) presentano un parametro di dispersione noto e pari a 1.

Il metodo della quasi-verosimiglianza permette di affrontare i problemi di sovradispersione e sottodispersione perchè, specificando un valore di ϕ derivante dalla stima, si va ad influenzare il valore di $var(Y_i)$ in modo da consentire una maggiore o minore variabilità rispetto a quella imposta dalla famiglia esponenziale.

La regressione binomiale caratterizzata da sovradispersione o sottodispersione può essere stimata in R selezionando `family=quasibinomial(link=-logit)` all'interno della funzione `glm`; tale opzione fa riferimento alla distribuzione Beta-Binomiale [Gelman, Hill 2006].

2.2.2 Distribuzione Beta-Binomiale

La distribuzione Beta-Binomiale è utilizzata per descrivere il numero di successi ottenuti in n esperimenti indipendenti dicotomici ed è caratterizzata dal fatto che la probabilità di successo π non è un parametro fisso, bensì si distribuisce come

una variabile casuale $Beta(a, b)$; siamo dunque di fronte ad una generalizzazione della distribuzione Binomiale.

Se $X \sim BeB(n, a, b)$ è una variabile casuale Beta-Binomiale di parametri n, a, b, allora, per $x \geq 0$, $P(X = x)$ può essere espressa in due modi equivalenti:

1. $$P(X = x) = C\binom{n}{x}\Gamma(a + x)\Gamma(b + n - x),$$

   dove C è una costante pari a

   $$C = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)\Gamma(a + b + n)}$$

   e $\Gamma()$ è la funzione Gamma, ovvero, per $\alpha > 0$,

   $$\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}\,dx = (\alpha - 1)\Gamma(\alpha - 1)$$

2. $$P(X = x) = \binom{n}{x}\frac{B(a + x, b + n - x)}{B(a, b)},$$

   dove $B()$ è la funzione Beta, ovvero, per $a > 0$ e $b > 0$,

   $$B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a, b)}.$$

Il valore atteso e la varianza sono così espressi:

$$E(X) = n\frac{a}{a + b}, \qquad var(X) = n\frac{ab}{(a + b)^2}\frac{a + b + n}{a + b + 1}.$$

Se poniamo $p = \frac{a}{a+b}$, otteniamo due formulazioni più vicine a quelle che caratterizzano la distribuzione Binomiale:

$$E(X) = np, \qquad var(X) = np(1 - p)\frac{a + b + n}{a + b + 1};$$

possiamo notare che, a parità di valore atteso e di n, la varianza della variabile casuale Beta-Binomiale è sempre superiore di quella della Binomiale.

2.3 GLM con funzioni del tempo come regressori

Focalizzeremo la nostra attenzione sulla modellazione della frazione di fumatori rilevata all'interno di gruppi di unità individuati tramite le seguenti variabili di stratificazione: classe d'età, sesso, livello di istruzione, regione. Decidiamo di affidarci al modello di regressione basato sulla distribuzione Beta-Binomiale e avente la funzione logistica come funzione legame.

Costruiremo dei modelli in funzione del tempo, ricorrendo alle componenti di trend e stagionalità, combinandole tra loro secondo uno schema additivo.

- Per esprimere il trend, la tendenza di fondo del fenomeno riferita ad un lungo periodo di tempo, ci affideremo a funzioni polinomiali del tempo, in sostituzione di altre funzioni più complesse ma non note. Le relative variabili esplicative saranno rappresentate da un vettore $t = 1, 2, \ldots, n$, con n pari al numero di periodi di tempo presi in considerazione nel modello, e dalle sue potenze, se necessarie.

- La componente stagionale, costituita dai movimenti del fenomeno nel corso dell'anno che tendono a ripetersi in maniera analoga nel medesimo periodo (nel nostro caso, mese) di anni successivi [Di Fonzo, Lisi 2005], verrà trattata in due diversi modi, che successivamente confronteremo:

  - tramite combinazioni di funzioni trigonometriche;
  - tramite variabili *dummy*.

  In generale, comunque, essa viene rappresentata mediante una funzione $h(t)$ necessariamente periodica, ossia tale per cui il suo valore all'istante t si riproduca esattamente ad intervalli costanti di lunghezza (periodo) s, con s pari a 12 nel caso di serie mensili: $h(t) = h(t + s) = h(t + 2s) = \ldots$.

2.3.1 Trend polinomiale e funzioni trigonometriche per la stagionalità

Proviamo a costruire dei modelli di regressione impostati sulla distribuzione Beta-Binomiale basandoci sulle seguenti funzioni del tempo:

- componente di trend:

$$T_t = \alpha_0 + \alpha_1 t + \cdots + \alpha_q t^q,$$

  dove il grado del polinomio q verrà scelto in base alla significatività dei parametri associati ai diversi regressori ed in base al test sulla devianza. Prestiamo attenzione al fatto che prendendo q troppo grande si rischia l'*overfitting*;

- componente stagionale, rappresentata dalla somma di m armoniche:

$$S_t = \sum_{i=1}^{m} A_i \cos\left(\frac{2\pi i}{s} t - \lambda_i\right);$$

  il generico addendo ha periodo pari a $\frac{s}{i}$, frequenza angolare $\omega_i = \frac{2\pi i}{s}$, ampiezza A_i e angolo di fase λ_i. Per dati mensili (s=12), la prima armonica (i=1) descrive un'onda cosinusoidale che completa il suo ciclo in 12 periodi di tempo, la seconda lo completa in 6 periodi e così via, la i-esima lo completa in $\frac{s}{i}$ periodi. Per serie storiche discrete, in cui t assume valori interi come nel nostro caso, si possono avere al massimo $m = [\frac{s}{2}]$ armoniche (6, se trattiamo dati mensili); di solito non è necessario considerare tutte le armoniche possibili, poichè già le prime generano dinamiche stagionali complesse.

  S_t può anche essere espressa con una funzione lineare dei parametri: sfruttando l'identità goniometrica $\cos(x - y) = \cos x \cos y + \sin x \sin y$, otteniamo

$$S_t = \sum_{i=1}^{m} (\beta_{i1} \cos \omega_i t + \beta_{i2} \sin \omega_i t),$$

con $\beta_{i1} = A_i \cos(\lambda_i)$ e $\beta_{i2} = A_i \sin(\lambda_i)$. Nel modello stimato con R utilizzeremo quest'ultima formulazione e andremo a verificare la significatività dei vari β_{i1} e β_{i2}.

Il predittore lineare, avrà dunque questa forma:

$$\eta_t = \alpha_0 + \alpha_1 t + \cdots + \alpha_q t^q + \sum_{i=1}^{m} (\beta_{i1} \cos \omega_i t + \beta_{i2} \sin \omega_i t), \qquad t = 1, \ldots, n$$

Prima di procedere con le stime dei modelli, esponiamo delle considerazioni che possono mettere in dubbio l'adeguatezza della scelta di questo schema di modellazione, in particolare il modo di trattare la componente stagionale:

1. i valori dei mesi di luglio ed agosto sono forzatamente uguali, come spiegato a pagina 3, di conseguenza è difficile pensare ad un andamento periodico regolare della serie;

2. nel corso dell'analisi descrittiva non avevamo notato andamenti periodici delle serie storiche che potessero far pensare alla presenza di una componente stagionale significativa.

Presentiamo i risultati ottenuti modellando la frazione di fumatori all'interno della classe d'età 18-34 anni (tabella 2.1); per la componente di trend abbiamo scelto un polinomio di grado 1, poichè aumentando il grado i relativi parametri non risultavano significativi, e abbiamo considerato m=3 armoniche nella costruzione della componente stagionale. Il predittore lineare è quindi

$$\eta_t = \alpha_0 + \alpha_1 t + \beta_{11} \cos \frac{\pi}{6} t + \beta_{12} \sin \frac{\pi}{6} t + \beta_{21} \cos \frac{\pi}{3} t + \beta_{22} \sin \frac{\pi}{3} t + \beta_{31} \cos \frac{\pi}{2} t + \beta_{32} \sin \frac{\pi}{2} t.$$

Se riteniamo significativo un parametro che presenta un *p-value*$\leq$0.05, notiamo che solamente l'intercetta, il parametro relativo alla componente di trend e β_{22} risultano tali. Ciò porta a concludere che, mentre la presenza di una tendenza di fondo (in questo caso decrescente, visto che α_1 è negativo) sembra essere confermata, non viene sostanzialmente riscontrata una componente stagionale. Il

parametro di dispersione stimato risulta pari a 0.5565: siamo dunque di fronte ad un caso di sottodispersione.

Tabella 2.1: Stime dei parametri del modello di regressione Beta-Binomiale (trend + stagionalità con funzioni trigonometriche) riferito alla frazione di fumatori all'interno della classe d'età 18-34 anni

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.5855	0.0139	-42.16	0.0000	***
α_1	-0.0026	0.0004	-5.98	0.0000	***
β_{11}	-0.0043	0.0097	-0.44	0.6608	
β_{12}	0.0036	0.0097	0.37	0.7146	
β_{21}	-0.0156	0.0097	-1.60	0.1159	
β_{22}	0.0234	0.0097	2.40	0.0204	*
β_{31}	-0.0098	0.0098	-1.01	0.3185	
β_{32}	0.0170	0.0096	1.76	0.0848	.

Anche ragionando sulle serie storiche di strati ottenuti tramite criteri diversi gli esiti non sono molto distanti da questi: raramente risultano significativi i parametri β_{i1} e β_{i2} e talvolta non lo sono nemmeno i parametri relativi al trend, cosicchè siamo costretti a propendere per un modello con la sola intercetta come parametro e nessun regressore temporale, come avremo modo di vedere successivamente.

2.3.2 Trend polinomiale e variabili *dummy* per la stagionalità

Passiamo alla costruzione di modelli di regressione impostati sulla distribuzione Beta-Binomiale e basati sulle seguenti funzioni del tempo:

- componente di trend polinomiale, come in precedenza;
- componente stagionale in cui la funzione periodica $h(t)$ ha la forma

$$h(t) = \sum_{j=1}^{s} \gamma_j d_{jt}, \qquad t = 1, 2, \ldots, n,$$

dove d_{jt} è una variabile *dummy* tale per cui

$$d_{jt} = \begin{cases} 1 & \text{nel periodo j-esimo dell'anno a cui appartiene t} \\ 0 & \text{altrimenti.} \end{cases}$$

Questa formulazione permette di cogliere l'effetto di ogni singolo mese sulla variabile Y.

Anche in questo caso, però, ci sentiamo di esprimere alcune perplessità riguardanti la componente stagionale.

1. Come già detto, i grafici presentati nel capitolo precedente non hanno evidenziato marcati comportamenti periodici del fenomeno.

2. Inserendo tutte le 12 variabili *dummy* nel modello, incappiamo nella multicollinearità: dobbiamo quindi escludere una mensilità, il cui effetto verrà inglobato nell'intercetta. Conseguentemente a ciò, i parametri stimati per gli 11 mesi inseriti come regressori saranno da considerarsi in riferimento al mese escluso; sarà invece difficoltoso capire l'effetto del mese escluso e la sua significatività ovvero capire se in quel mese si registra o meno un comportamento pressochè analogo nel corso degli anni.

3. Poichè non vi è una regola che stabilisce quale mensilità escludere, la scelta è arbitraria, ma non ininfluente: cambiando mese "di riferimento" variano le stime dei parametri che si riferiscono ai mesi inclusi come regressori ed anche le loro significatività.

4. Potrebbero risultare significativi solo i parametri di poche variabili *dummy* e sorgerebbero quindi i dubbi sul modello da adottare: tenere in blocco tutte le 11 variabili, sebbene la maggior parte non risulti significativa? Abbandonare l'idea della presenza di una componente stagionale?

Una soluzione potrebbe essere quella di inserire una alla volta le mensilità, valutarne la significatività e provare poi a combinarle secondo determinati criteri.

Come in un problema di selezione delle variabili, sviluppiamo la costruzione di modelli secondo due approcci distinti: prima esamineremo l'approccio *backward*, in cui si parte da un modello avente tutte le variabili d'interesse all'interno per giungere, tramite valutazione della significatività, ad una scrematura di esse; in seguito ci affideremo all'approccio *forward*, in cui si valuta un regressore per volta, si sceglie quello che conduce al miglior adattamento dei dati, si esaminano i restanti aggiungendoli uno alla volta al modello e si include quello che apporta più informazione.

2.3.2.1 Approccio *backward*

Consideriamo come Y_i, come fatto in precedenza, la frazione di fumatori rilevata tra gli intervistati appartenenti alla classe d'età 18-34 anni. Optiamo per una componente di trend polinomiale di grado 1 ed inseriamo le 11 variabili *dummy*, escludendo il mese di gennaio, che sarà dunque quello di riferimento. La tabella 2.2 riporta i valori stimati dei parametri e la loro significatività: solo l'intercetta e la componente di trend risultano significativamente diverse da zero. Possiamo interpretare i risultati ottenuti in questo modo:

- il fenomeno segue un andamento lineare decrescente;
- negli 11 mesi inseriti come regressori, non si registrano significative differenze rispetto alla media e rispetto all'andamento del mese di gennaio.

Cambiando mese di riferimento ed escludendo, per esempio, febbraio, dal modello, i risultati cambiano: il parametro che si riferisce ad aprile risulta significativo e la sua stima ha segno negativo, il che vuol dire che nel mese di aprile si registra una frazione di fumatori significativamente più bassa rispetto alla media e rispetto al mese di febbraio (tabella 2.3).

Effettuiamo un test d'ipotesi per verificare se la devianza spiegata dal modello contenente trend e variabili *dummy* risulta significativamente diversa da quella

Tabella 2.2: Stime dei parametri del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori all'interno della classe d'età 18-34 anni. Approccio *backward*, versione 1

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.586099	0.027834	-21.057	< 2e-16	***
α_1	-0.002617	0.000441	-5.934	4.94e-07	***
d_2	0.033776	0.036249 0	.932	0.357	
d_3	0.004491	0.036750	0.122	0.903	
d_4	-0.060234	0.036000	-1.673	0.102	
d_5	-0.034786	0.035899	-0.969	0.338	
d_6	0.050046	0.034620	1.446	0.156	
d_7	0.016318	0.033490	0.487	0.629	
d_8	0.018936	0.033499	0.565	0.575	
d_9	-0.021473	0.034492	-0.623	0.537	
d_{10}	-0.004044	0.034130	-0.118	0.906	
d_{11}	-0.008054	0.034560	-0.233	0.817	
d_{12}	0.015102	0.034963	0.432	0.668	

spiegata dal modello con la sola componente di trend, a fronte di un consistente aumento del numero di parametri, con conseguente diminuzione dei gradi di libertà. Ci basiamo sulla distribuzione di Fisher, poichè ϕ non è noto, ed otteniamo che non vi è un significativo miglioramento del modello (tabella 2.4): la devianza residua è più bassa nel modello contenente le variabili *dummy*, ma questa evidentemente è solamente una naturale conseguenza dell'aumento del numero di parametri e non è da considerarsi un beneficio significativo dovuto all'introduzione di quegli specifici regressori.

Riteniamo che l'approccio *backward* sia poco produttivo, poichè non riesce ad offrire una soluzione alle perplessità in precedenza esposte.

2.3.2.2 Approccio *forward*

Dopo aver individuato il grado del polinomio della componente di trend più adatto, procediamo all'inserimento di una variabile *dummy* alla volta, per verificare la presenza o meno di qualche "effetto di mese" nella serie presa in considerazione. Come già detto, abbiamo forti dubbi sulla presenza di una compo-

Tabella 2.3: Stime dei parametri del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori all'interno della classe d'età 18-34 anni. Approccio *backward*, versione 2

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.552323	0.028244	-19.555	< 2e-16	***
α_1	-0.002617	0.000441	-5.934	4.94e-07	***
d_1	-0.033776	0.036249	-0.932	0.3568	
d_3	-0.029285	0.036963	-0.792	0.4326	
d_4	-0.094010	0.036215	-2.596	0.0129	*
d_5	-0.068562	0.036112	-1.899	0.0645	.
d_6	0.016270	0.034855	0.467	0.6431	
d_7	-0.017458	0.033730	-0.518	0.6075	
d_8	-0.014840	0.033734	-0.440	0.6622	
d_9	-0.055249	0.034719	-1.591	0.1190	
d_{10}	-0.037820	0.034358	-1.101	0.2773	
d_{11}	-0.041830	0.034780	-1.203	0.2358	
d_{12}	-0.018674	0.035176	-0.531	0.5983	

Tabella 2.4: Test ANOVA applicato ai modelli "trend" e "trend+*dummy*", riferiti alla frazione di fumatori all'interno della classe d'età 18-34 anni

	Gdl	Dev. res.	Diff. gdl	Diff. dev.	F	Pr(>F)	Signif.
trend	53	33.18					
trend+*dummy*	42	23.83	11	9.34	1.50	0.1691	

nente stagionale marcata, quindi arrivati a questo punto ci limitiamo a valutare eventuali comportamenti particolari riscontrati di mese in mese.

Ci concentreremo sulla frazione di fumatori riscontrata tra gli intervistati presi complessivamente e suddivisi per classi d'età, sesso, livello d'istruzione, regione.

Totale intervistati

Il modello stimato per rappresentare la frazione di fumatori registrata al interno di tutto il campione prevede una componente di trend lineare: la stima del parametro α_1 risulta essere negativa, ad indicare un andamento decrescente del fenomeno (tabella 2.5). Nessuna variabile *dummy* è significativa: in nessun mese risulta esserci un comportamento significativamente diverso rispetto a quello

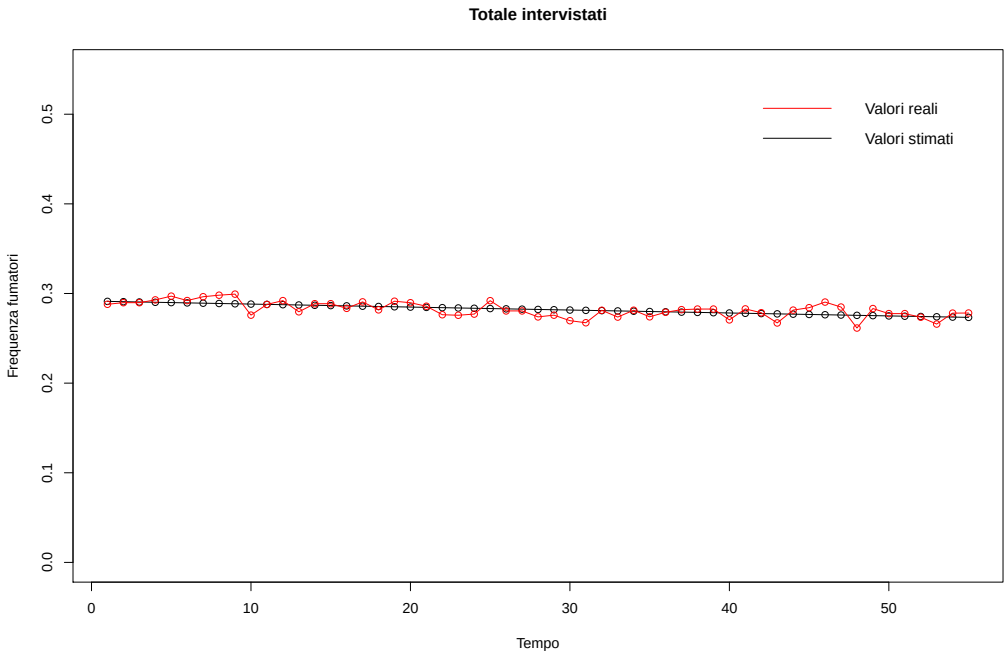
descritto dalla retta individuata dalla componente di trend. Il grafico 2.1 mette a confronto la serie dei dati osservati e la serie dei dati stimati. Si notano delle oscillazioni che il modello non riesce a cogliere, che non sono però riconducibili ad una qualche componente stagionale, bensì ad una componente d'errore. Il modello riesce a spiegare circa il 38% di variabilità, a fronte della presenza di un unico regressore.

Tabella 2.5: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori complessiva. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.8878949	0.0091516	-97.02	< 2e-16	***
α_1	-0.0016309	0.0002856	-5.71	5.22e-07	***

	Valore	Gdl
Devianza nulla	65.689	54
Devianza residua	40.693	53

Figura 2.1: Frequenza fumatori - totale intervistati; serie osservata VS serie stimata



Classe d'età

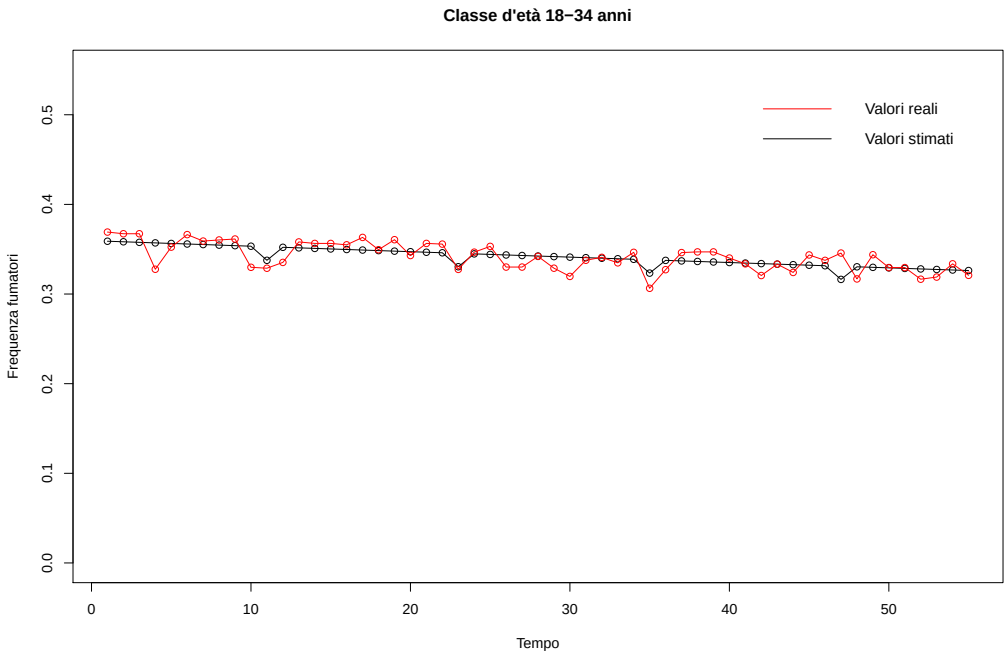
La percentuale di fumatori tra gli individui appartenenti alla classe d'età 18-34 anni è modellata da una componente di trend lineare avente coefficiente negativo e presenta un "effetto di mese" riferito ad aprile, in cui si registra un picco verso il basso (meno fumatori). In dettaglio, si vedano la tabella 2.6 ed il grafico 2.2. La variabilità spiegata dal modello supera il 46%.

Tabella 2.6: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori nella classe d'età 18-34. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.5773943	0.0140547	-41.082	< 2e-16	***
α_1	-0.0026895	0.0004387	-6.131	1.19e-07	***
d_4	-0.0668784	0.0263786	-2.535	0.0143	*

	Valore	Gdl
Devianza nulla	54.749	54
Devianza residua	29.510	52

Figura 2.2: Frequenza fumatori - classe d'età 18-34; serie osservata VS serie stimata



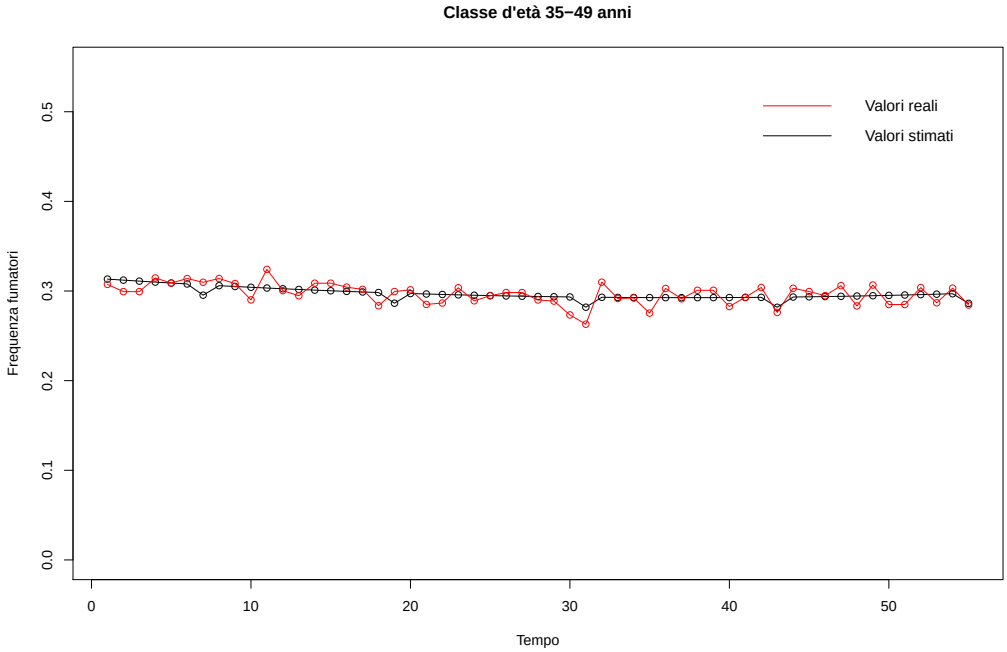
Per quanto riguarda la frazione riscontrata tra coloro che hanno un'età compresa tra i 35 e i 49 anni, il modello individuato è composto da una componente di trend di tipo quadratico (parabola con leggera concavità verso l'alto, ad indicare un andamento prima decrescente e poi crescente) e dalla variabile *dummy* che fa riferimento al mese di dicembre, in cui si registrano dei picchi verso il basso (tabella 2.7 e grafico 2.3). La variabilità spiegata si aggira intorno al 31%.

Tabella 2.7: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori nella classe d'età 35-49. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-7.793e-01	2.054e-02	-37.935	<2e-16	***
α_1	-5.571e-03	1.670e-03	-3.335	0.0016	**
α_2	7.489e-05	2.882e-05	2.598	0.0122	*
d_4	-5.485e-02	2.312e-02	-2.373	0.0215	*

	Valore	Gdl
Devianza nulla	41.810	54
Devianza residua	28.582	51

Figura 2.3: Frequenza fumatori - classe d'età 35-49; serie osservata VS serie stimata



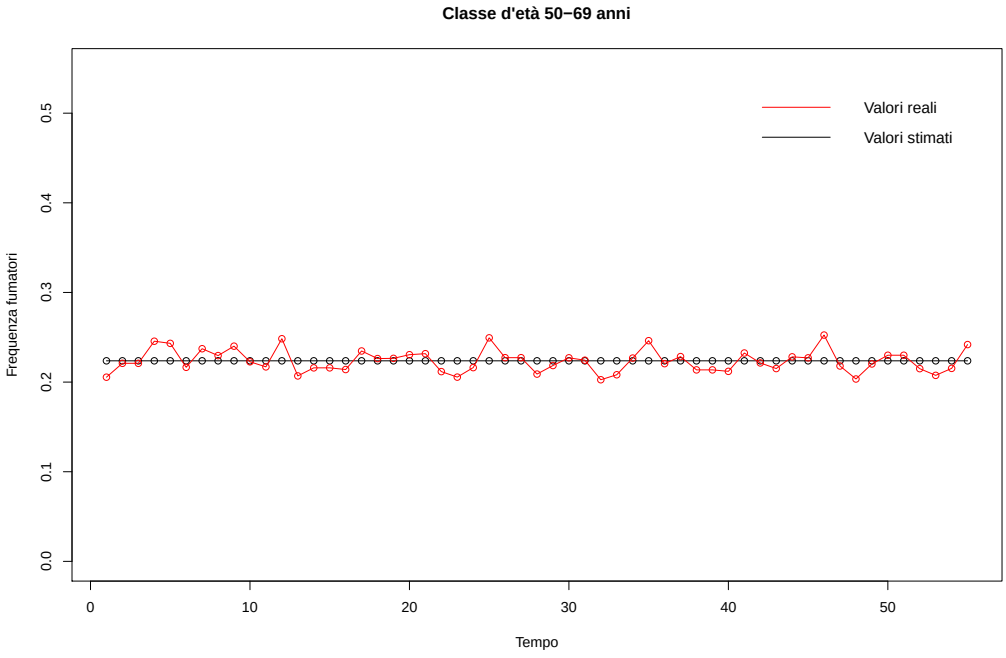
Il modello individuato per la classe d'età 50-69 non presenta alcun predittore: il trend risulta quindi costante e non vi sono "effetti di mese". La variabilità dei valori osservati è riconducibile totalmente ad una componente di errore che non riusciamo significativamente a modellare (tabella 2.8 e grafico 2.4).

Tabella 2.8: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori nella classe d'età 50-69. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.244037	0.009704	-128.2	<2e-16	***

	Valore	Gdl
Devianza nulla	61.458	54
Devianza residua	61.458	54

Figura 2.4: Frequenza fumatori - classe d'età 50-69; serie osservata VS serie stimata



Sesso

La percentuale di fumatori tra gli individui di sesso maschile può essere modellata da una retta avente coefficiente angolare negativo e che rappresenta la

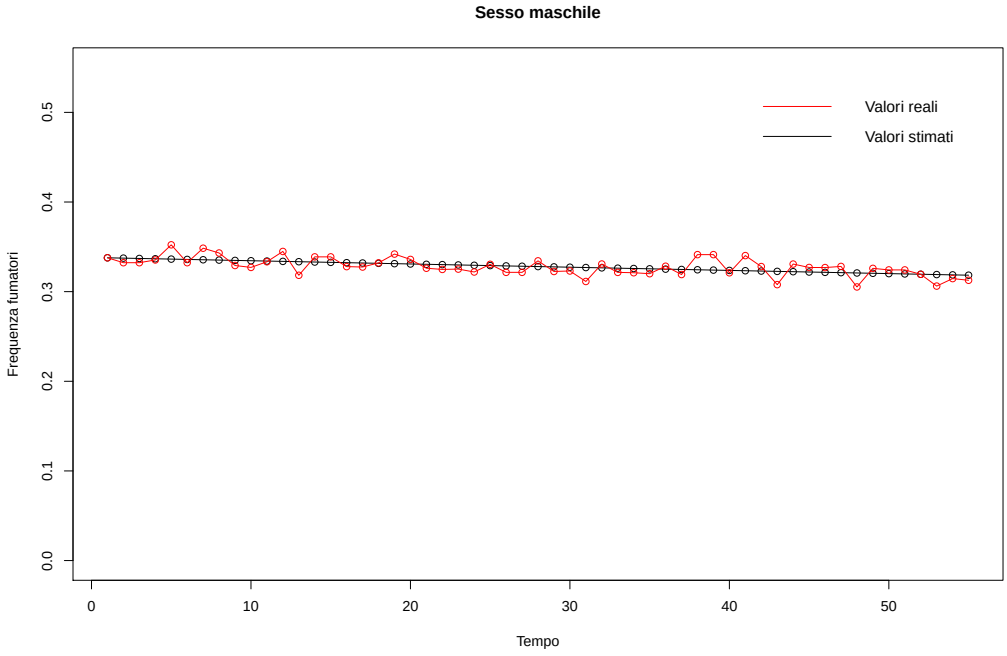
componente polinomiale di trend, che riesce a spiegare circa il 31% della variabilità. Si vedano la tabella 2.9 ed il grafico 2.5 per maggiori dettagli.

Tabella 2.9: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori di sesso maschile. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.6716451	0.0105907	-63.419	<2e-16	***
α_1	-0.0016305	0.0003302	-4.938	8.26e-06	***

	Valore	Gdl
Devianza nulla	42.376	54
Devianza residua	29.014	53

Figura 2.5: Frequenza fumatori - sesso maschile; serie osservata VS serie stimata



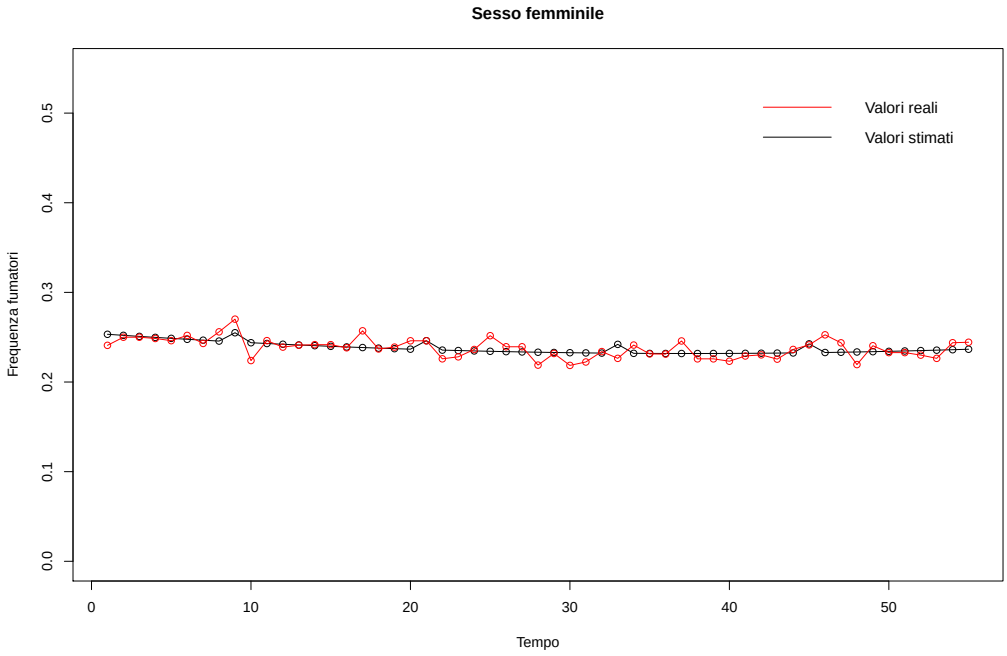
La frazione di fumatrici viene modellata da un polinomio di secondo grado per quanto riguarda il trend, che quindi assume la forma di una parabola con lieve concavità verso l'alto, e dalla variabile *dummy* riferita al mese di febbraio, che evidenzia una percentuale più alta in quel mese (tabella 2.10 e grafico 2.6). Il modello spiega il 35% di variabilità.

Tabella 2.10: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori di sesso femminile. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.075e+00	2.110e-02	-50.934	<2e-16	***
α_1	-6.564e-03	1.736e-03	-3.781	0.000411	***
α_2	8.738e-05	3.000e-05	2.912	0.005310	**
d_2	5.464e-02	2.616e-02	2.089	0.041761	*

	Valore	Gdl
Devianza nulla	60.925	54
Devianza residua	39.561	51

Figura 2.6: Frequenza fumatori - sesso femminile; serie osservata VS serie stimata



Livello d'istruzione

Per modellare la percentuale di fumatori riscontrata tra coloro che non possiedono alcun titolo di studio oppure hanno conseguito la licenza elementare, ci sembra adeguata una componente di trend polinomiale di terzo grado, cui non aggiungiamo alcuna variabile *dummy* perchè nessuna di esse è risultata significa-

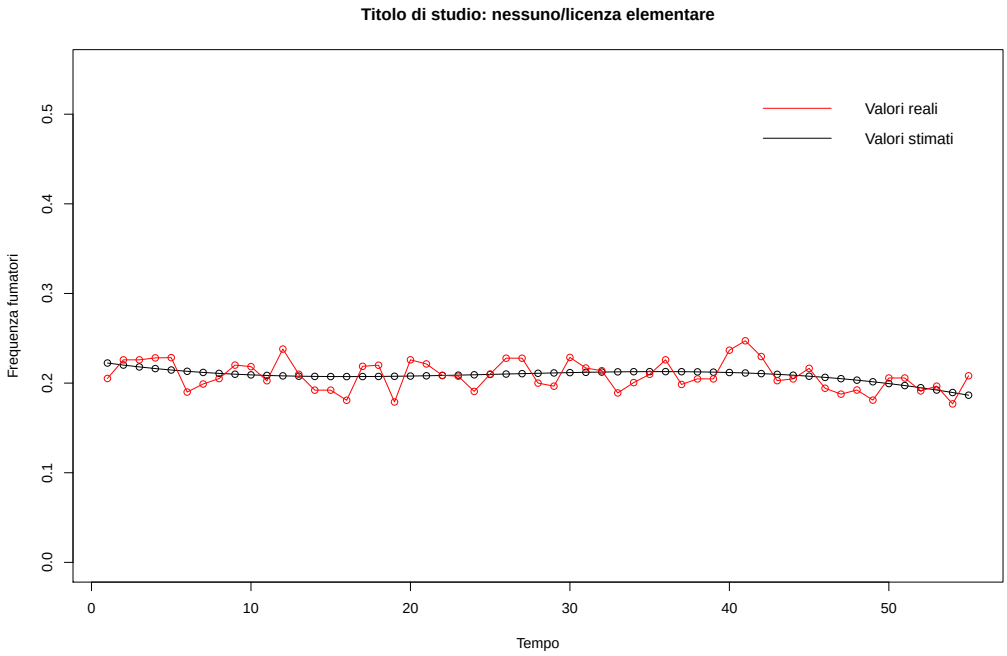
tiva dopo averla inserita nel modello secondo l'approccio *forward* (tabella 2.11 e grafico 2.7). La variabilità spiegata sfiora il 15%, percentuale piuttosto bassa.

Tabella 2.11: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori con nessun titolo di studio/licenza elementare. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.237e+00	5.114e-02	-24.190	<2e-16	***
α_1	-1.532e-02	8.076e-03	-1.897	0.0634	.
α_2	6.932e-04	3.430e-04	2.021	0.0486	*
α_3	-8.957e-06	4.120e-06	-2.174	0.0344	*

	Valore	Gdl
Devianza nulla	35.069	54
Devianza residua	29.962	51

Figura 2.7: Frequenza fumatori - nessun titolo di studio/licenza elementare; serie osservata VS serie stimata



La serie che rappresenta la frazione di fumatori tra coloro che possiedono la licenza di scuola media inferiore è modellata da una componente di trend lineare decrescente e da un "effetto di mese", riferito a maggio, in cui si individua un

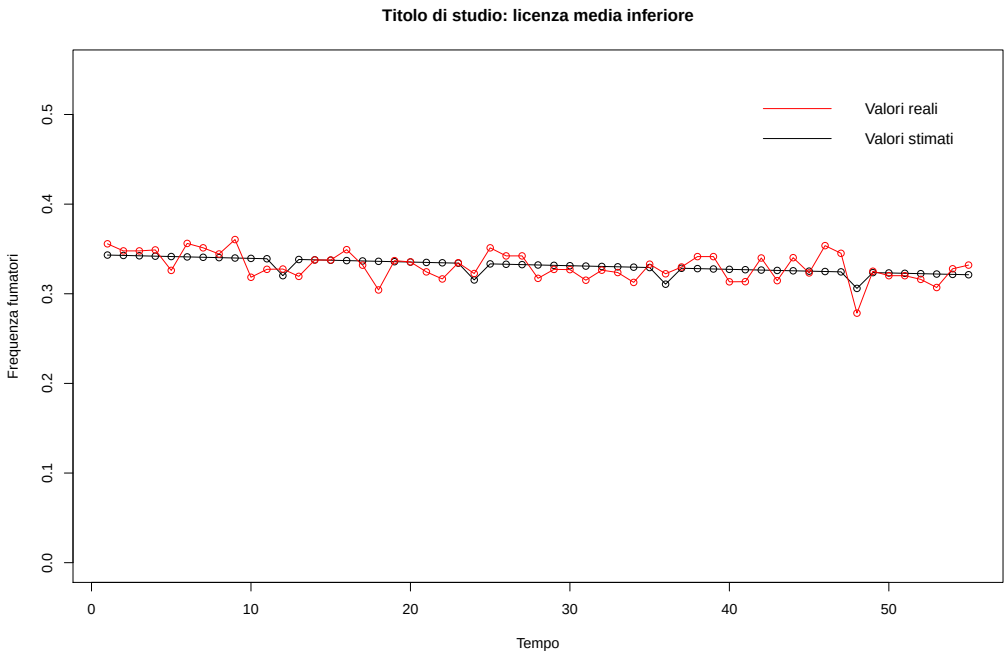
picco verso il basso; la tabella 2.12 ed il grafico 2.8 ci illustrano quanto appena detto. In questo caso il modello riesce a spiegare circa il 28% della variabilità dei dati.

Tabella 2.12: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori con licenza scuola media inferiore. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.6472599	0.0166454	-38.885	<2e-16	***
α_1	-0.0018347	0.0005168	-3.550	0.000827	***
d_5	-0.0835832	0.0307610	-2.717	0.008921	**

	Valore	Gdl
Devianza nulla	62.078	54
Devianza residua	44.504	52

Figura 2.8: Frequenza fumatori - licenza scuola media inferiore; serie osservata VS serie stimata



Come nel primo livello di istruzione analizzato, anche per rappresentare la frazione di fumatori tra gli individui con diploma di scuola superiore scegliamo

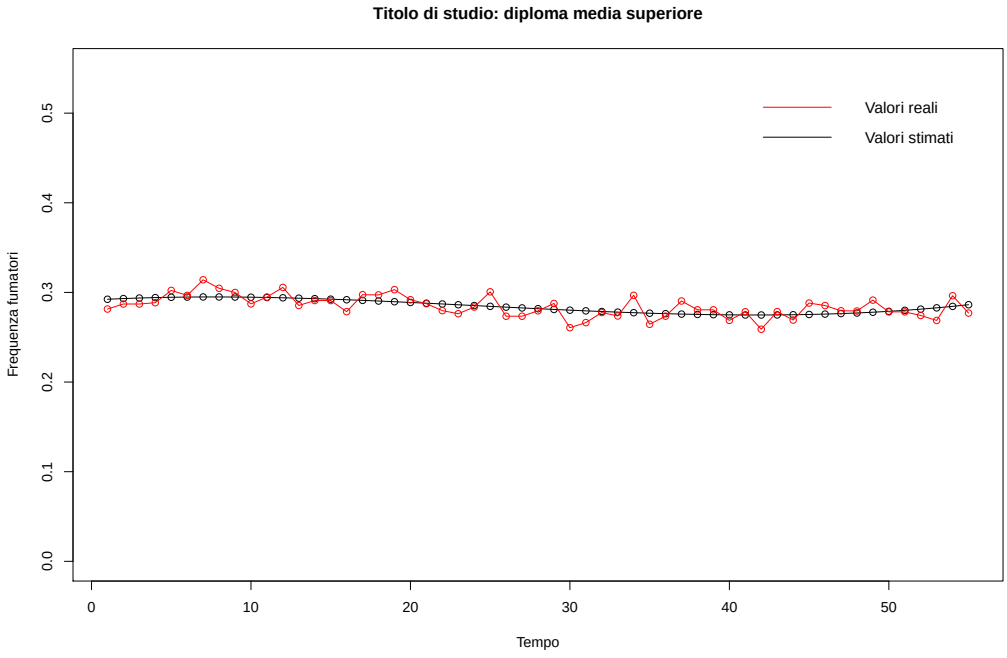
una componente di trend polinomiale di grado 3 (tabella 2.13 e grafico 2.9); in questo caso essa riesce a spiegare circa il 38% di variabilità.

Tabella 2.13: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori con diploma scuola media superiore. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-8.882e-01	2.856e-02	-31.105	<2e-16	***
α_1	4.718e-03	4.292e-03	1.099	0.2768	
α_2	-3.679e-04	1.760e-04	-2.091	0.0416	*
α_3	4.977e-06	2.060e-06	2.416	0.0193	*

	Valore	Gdl
Devianza nulla	56.125	54
Devianza residua	36.665	51

Figura 2.9: Frequenza fumatori - diploma scuola media superiore; serie osservata VS serie stimata



Per quanto riguarda chi possiede una laurea o un diploma universitario, la frazione di fumatori riscontrata tra essi viene modellata attraverso un trend li-

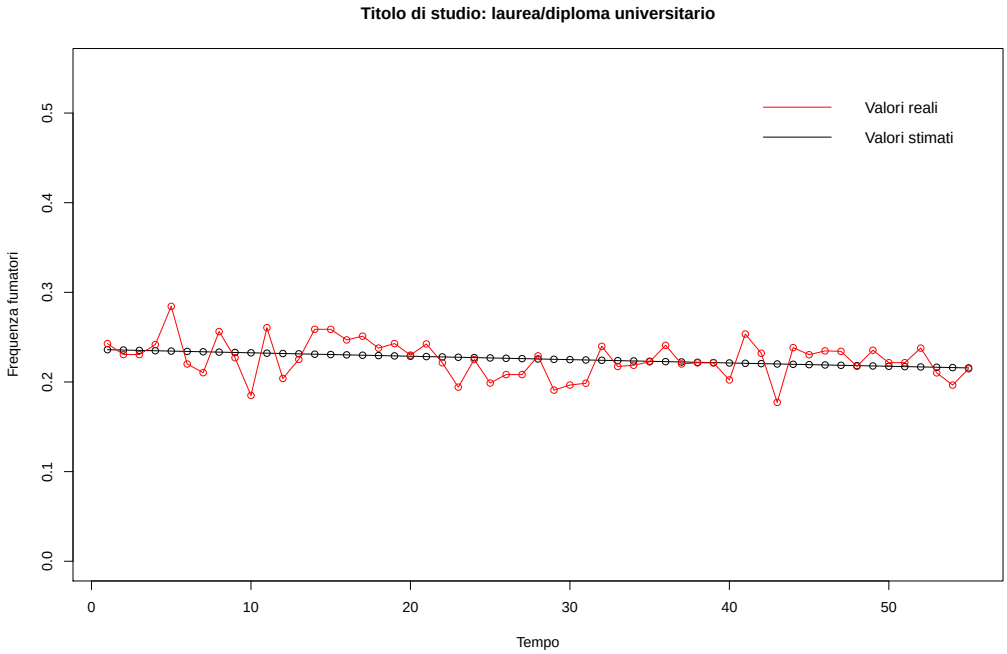
neare decrescente che però riesce a spiegare solo poco più dell'8% di variabilità (tabella 2.14 e grafico 2.10); nessuna variabile *dummy* risulta essere significativa.

Tabella 2.14: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori con diploma scuola media superiore. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.1724978	0.0321414	-36.479	<2e-16	***
α_1	-0.0021530	0.0009856	-2.185	0.0334	*

	Valore	Gdl
Devianza nulla	60.029	54
Devianza residua	55.105	53

Figura 2.10: Frequenza fumatori - laurea/diploma universitario; serie osservata VS serie stimata



Regione

Se consideriamo come variabile di stratificazione la regione di appartenenza, dobbiamo fare attenzione al fatto che in alcuni mesi, in determinate regioni (Lombardia, Lazio, Molise, Puglia, Basilicata, Calabria, Sardegna), non sono stati

rilevati i dati oppure la numerosità campionaria risulta molto più bassa di quella riscontrata nelle restanti mensilità, così da indurre ad escludere i relativi dati dall'analisi. Di conseguenza, il modello sarà stimato sulla base di un numero minore di osservazioni. Esso comunque ci risulterà utile anche per ottenere una stima della frequenza dei fumatori prevista nelle mensilità mancanti di osservazioni.

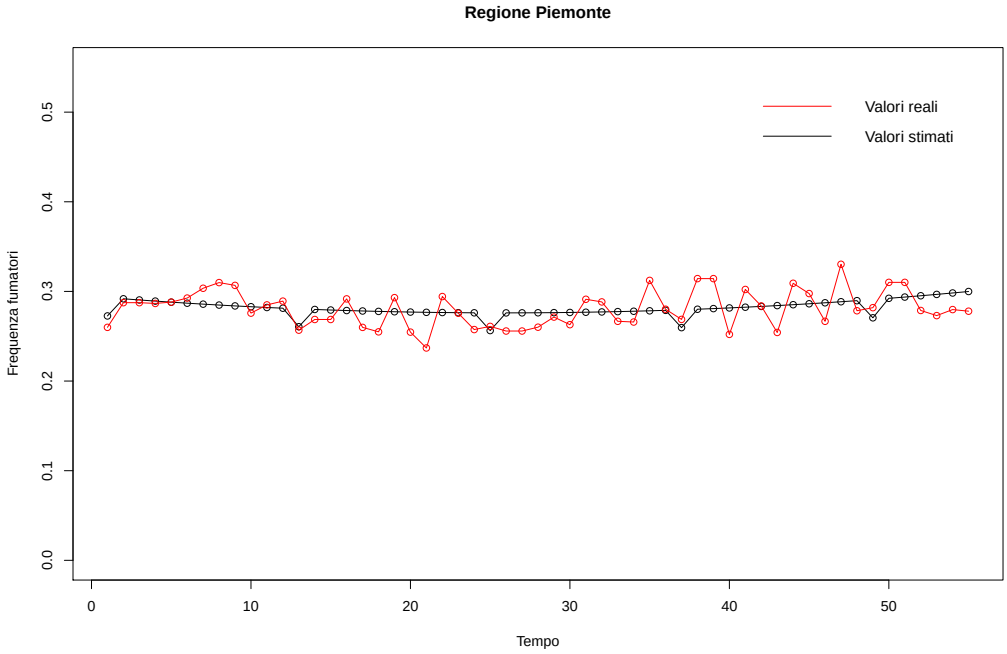
Il modello che riteniamo più adeguato per il fenomeno registrato tra gli abitanti del Piemonte presenta un trend quadratico, che individua un andamento prima decrescente e poi crescente, ed un "effetto di mese" riferito a giugno che sta ad indicare una diminuzione della frazione di fumatori in quella mensilità, ogni anno (tabella 2.15 e grafico 2.11). La variabilità spiegata supera di poco il 17%.

Tabella 2.15: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Piemonte. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-8.732e-01	3.550e-02	-24.598	<2e-16	***
α_1	-7.045e-03	3.131e-03	-2.250	0.0288	*
α_2	1.365e-04	5.641e-05	2.420	0.0191	*
d_6	-1.011e-01	4.522e-02	-2.235	0.0298	*

	Valore	Gdl
Devianza nulla	44.049	54
Devianza residua	36.550	51

Figura 2.11: Frequenza fumatori - regione Piemonte; serie osservata VS serie stimata



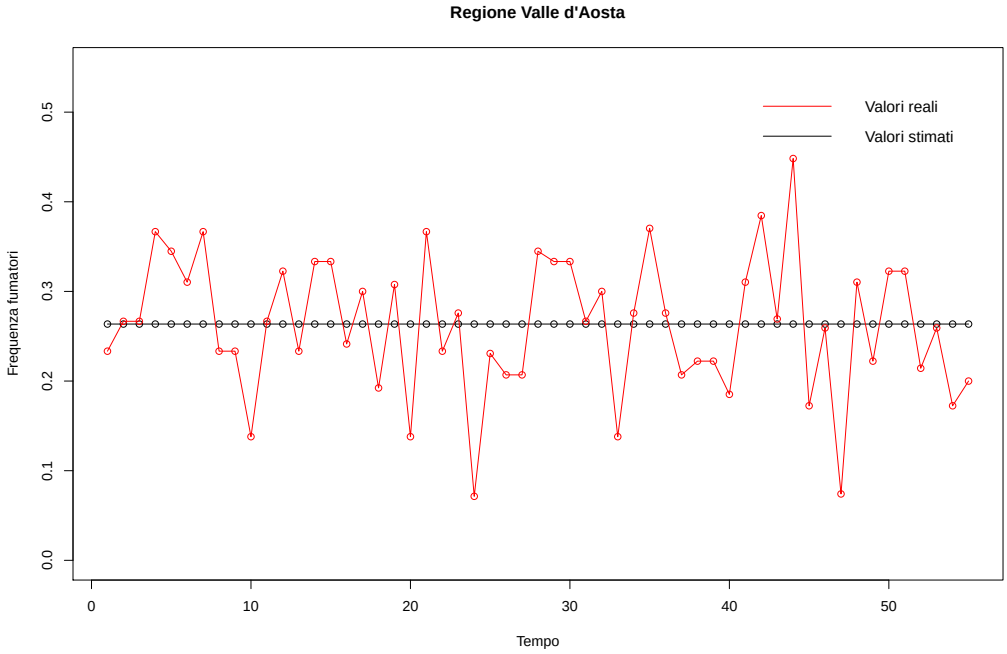
Per la Valle d'Aosta non riusciamo ad individuare alcun parametro significativo, nè per quanto riguarda la componente di trend polinomiale, nè per la parte "stagionale"; questo può essere dovuto forse al fatto che i dati presentano una forte variabilità, a causa della bassa numerosità dei campioni rilevati mensilmente. Ci affidiamo al modello nullo, in cui compare solo l'intercetta (tabella 2.16 e grafico 2.12).

Tabella 2.16: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Valle d'Aosta. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.02739	0.05342	-19.23	<2e-16	***

	Valore	Gdl
Devianza nulla	50.233	54
Devianza residua	50.233	54

Figura 2.12: Frequenza fumatori - regione Valle d'Aosta; serie osservata VS serie stimata



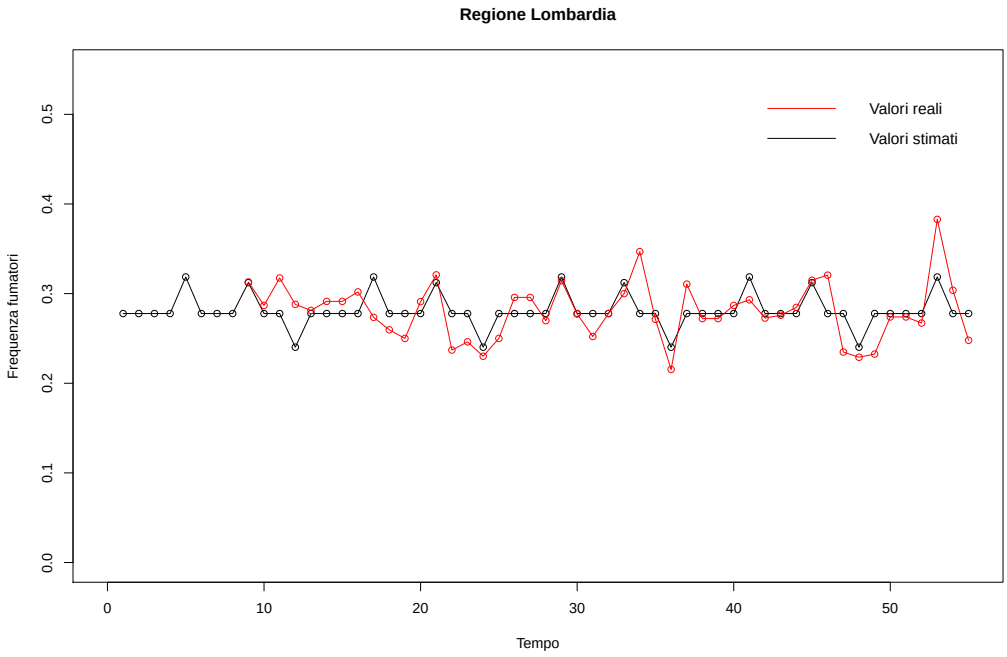
La Lombardia presenta otto mensilità mancanti: da giugno 2007 a gennaio 2008 compresi, non sono disponibili i dati. Il modello è quindi stimato sulla base di 47 valori, non di 55 come nei casi precedentemente esaminati. Il trend risulta essere costante e si individuano tre variabili *dummy* significative, che si riferiscono ai mesi di febbraio, maggio e ottobre e che vengono inserite nel modello (tabella 2.17 e grafico 2.13). In particolare, nei mesi di febbraio ed ottobre si registrano dei picchi verso l'alto, mentre nei mesi di maggio la frazione di fumatori tende ad essere più bassa rispetto alla media. Queste variabili riescono a spiegare quasi il 34% di variabilità.

Tabella 2.17: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Lombardia. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.95557	0.02279	-41.923	<2e-16	***
d_2	0.16600	0.06866	2.418	0.01992	*
d_5	-0.19582	0.07450	-2.628	0.01185	*
d_10	0.19520	0.07134	2.736	0.00899	**

	Valore	Gdl
Devianza nulla	30.363	46
Devianza residua	20.083	43

Figura 2.13: Frequenza fumatori - regione Lombardia; serie osservata VS serie stimata



La serie che si riferisce ai fumatori della provincia autonoma di Bolzano presenta oscillazioni piuttosto ampie: la causa, come nel caso della Valle d'Aosta, può essere ricercata nella bassa numerosità campionaria. Questa oscillazioni non sono però riconducibili ad "effetti di mese" significativi ed il modello stimato per questa serie di dati presenta solamente una componente di trend lineare de-

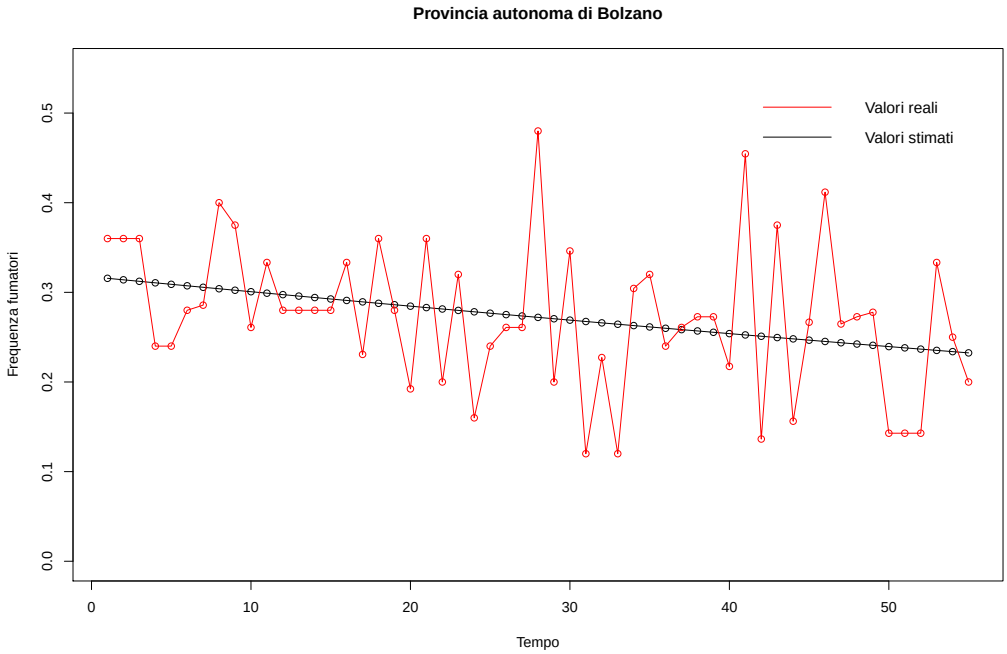
crescente che riesce a spiegare poco meno del 10% di variabilità (tabella 2.18 e grafico 2.14).

Tabella 2.18: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della provincia autonoma di Bolzano. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.766017	0.110752	-6.916	6.15e-09	***
α_1	-0.007789	0.003369	-2.312	0.0247	*

	Valore	Gdl
Devianza nulla	51.333	54
Devianza residua	46.637	53

Figura 2.14: Frequenza fumatori - provincia autonoma di Bolzano; serie osservata VS serie stimata



Anche la serie che rappresenta la frequenza dei fumatori tra gli abitanti della provincia autonoma di Trento è parecchio irregolare e la numerosità campionaria è bassa. Il trend è costante e viene riscontrato un significativo "effetto di mese" che si riferisce a dicembre, mese in cui la percentuale risulta essere più alta della

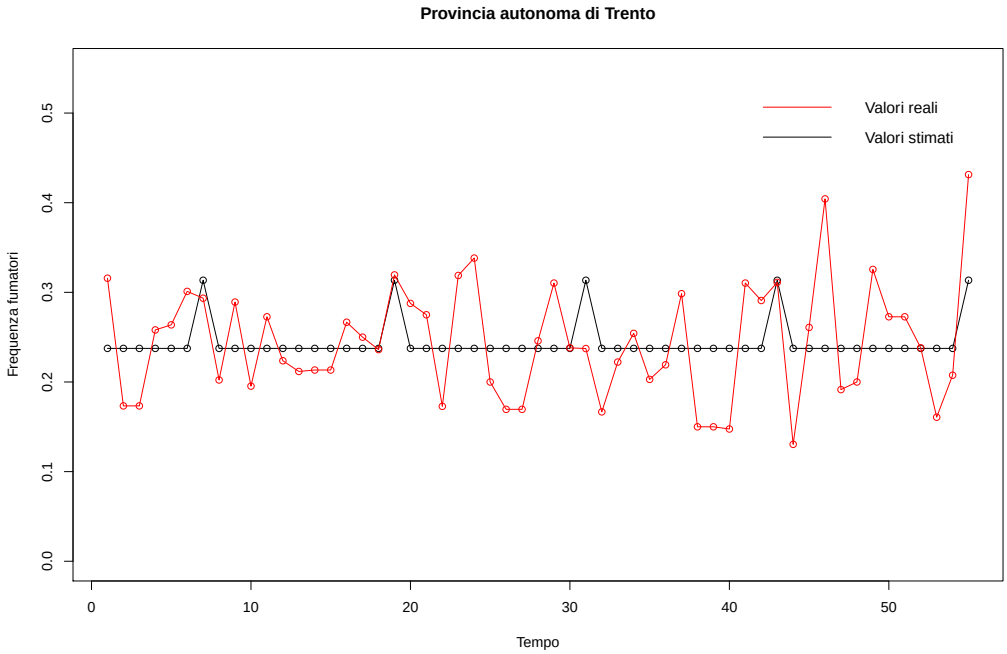
media. Questa variabile *dummy* spiega circa il 12% della variabilità totale (tabella 2.19 e grafico 2.15).

Tabella 2.19: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della provincia autonoma di Trento. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.16656	0.04458	-26.17	< 2e-16	***
$d_1 2$	0.38266	0.13915	2.75	0.00813	**

	Valore	Gdl
Devianza nulla	71.948	54
Devianza residua	63.292	53

Figura 2.15: Frequenza fumatori - provincia autonoma di Trento; serie osservata VS serie stimata



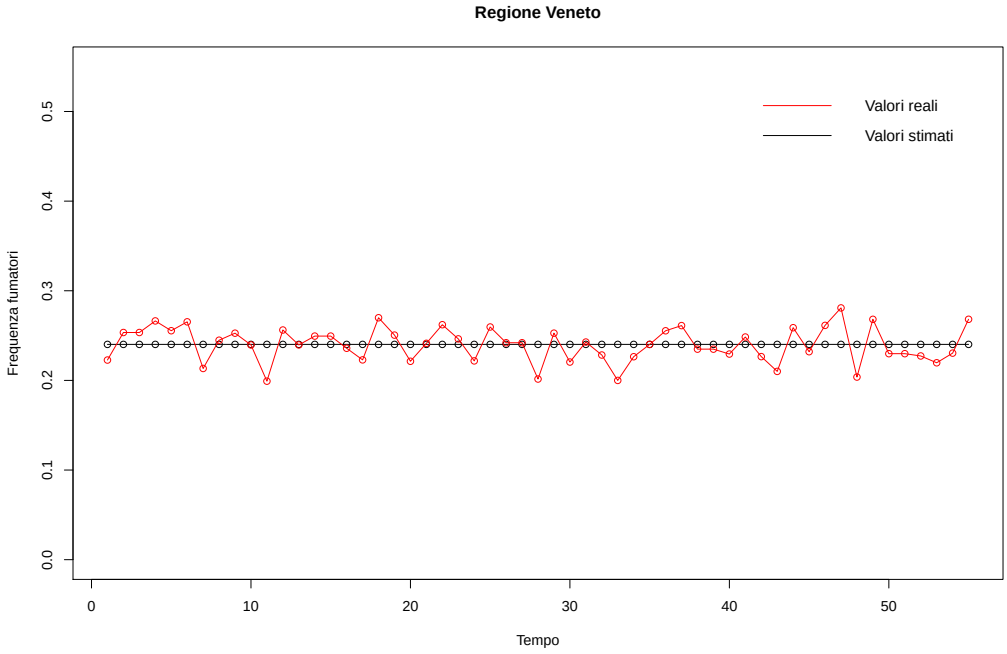
Per quanto riguarda i dati relativi alla regione Veneto, non risultano significativi nè i parametri che si riferiscono alla componente di trend nè quella che si riferiscono alle variabili *dummy* (tabella 2.20 e grafico 2.16). Il modello è composto solamente dall'intercetta e le oscillazioni rientrano nella componente d'errore.

Tabella 2.20: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Veneto. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.15167	0.01422	-81.01	< 2e-16	***

	Valore	Gdl
Devianza nulla	51.156	54
Devianza residua	51.156	54

Figura 2.16: Frequenza fumatori - regione Veneto; serie osservata VS serie stimata



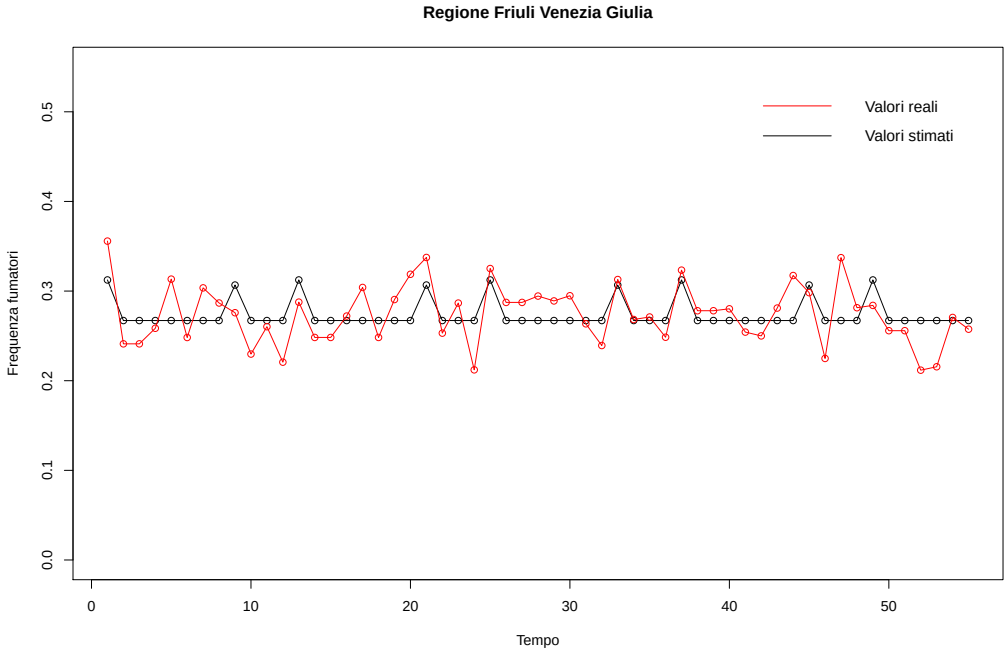
Per modellare i dati relativi agli abitanti del Friuli Venezia Giulia optiamo per un trend costante e per due variabili *dummy*, relative a febbraio e giugno: in questi due mesi si registra una frazione di fumatori significativamente più alta rispetto alla media, soprattutto nel mese di giugno (tabella 2.21 e grafico 2.17). La variabilità spiegata è quasi del 23%.

Tabella 2.21: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Friuli Venezia Giulia. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.00924	0.02175	-46.402	< 2e-16	***
d_2	0.19379	0.07414	2.614	0.01169	*
d_6	0.22040	0.06867	3.209	0.00228	**

	Valore	Gdl
Devianza nulla	46.080	54
Devianza residua	35.552	52

Figura 2.17: Frequenza fumatori - regione Friuli Venezia Giulia; serie osservata VS serie stimata



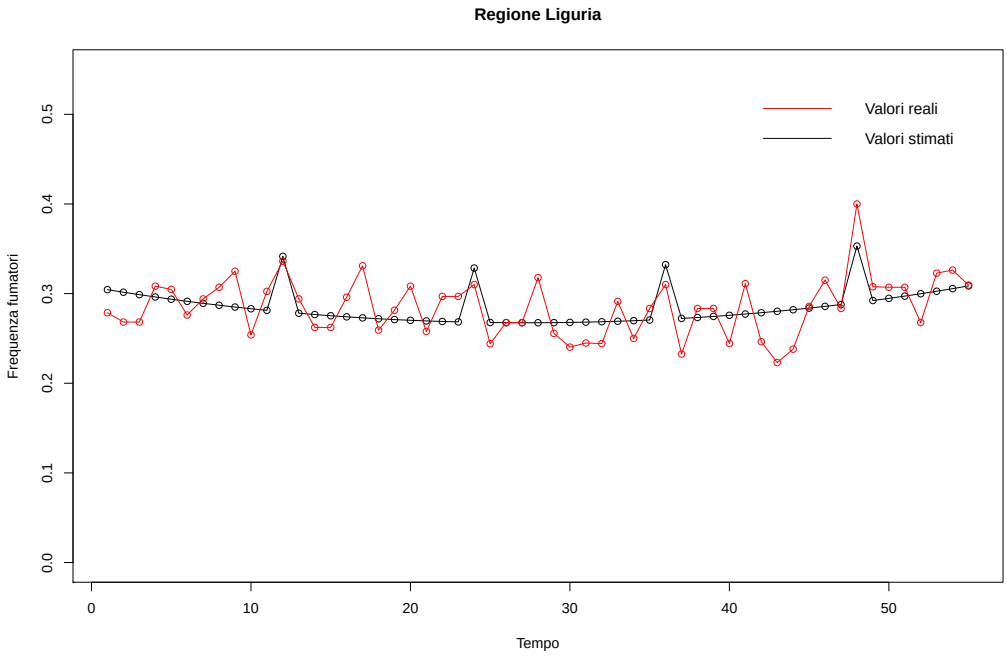
Trend quadratico e variabile *dummy* riferita al mese di maggio sono i predittori scelti per modellare la serie dei dati della Liguria: il grafico presenta una parabola con concavità verso l'alto e dei picchi verso l'alto in corrispondenza dei mesi di maggio. I tre regressori inseriti spiegano quasi il 35% di variabilità (tabella 2.22 e grafico 2.18).

Tabella 2.22: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Liguria. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.8122739	0.0558418	-14.546	< 2e-16	***
α_1	-0.0142940	0.0046142	-3.098	0.00317	**
α_2	0.0002617	0.0000796	3.288	0.00183	**
d_5	0.2897954	0.0674114	4.299	7.74e-05	***

	Valore	Gdl
Devianza nulla	36.026	54
Devianza residua	23.639	51

Figura 2.18: Frequenza fumatori - regione Liguria; serie osservata VS serie stimata



La serie che si riferisce all'Emilia Romagna può essere modellata da una componente di trend costante e dalle variabili *dummy* che si riferiscono ai mesi di aprile, in cui si registra una frazione di fumatori significativamente più bassa della media, e di giugno, in cui invece la percentuale di fumatori è più alta (tabella 2.23 e grafico 2.19). Queste due "effetti di mese" spiegano quasi il 19% di

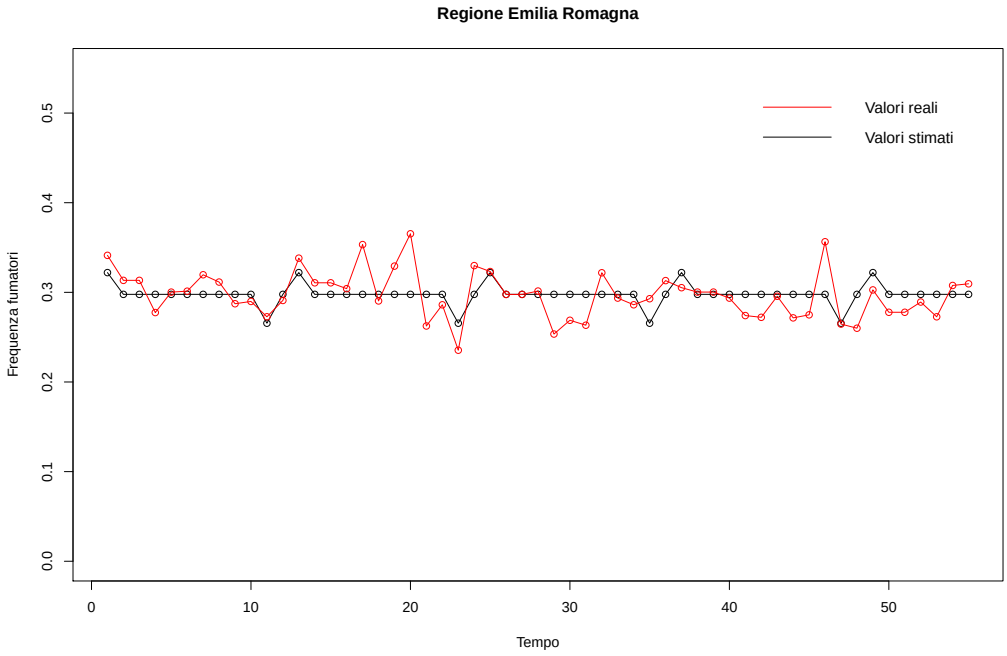
variabilità.

Tabella 2.23: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Emilia Romagna. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.85797	0.01713	-50.087	< 2e-16	***
d_4	-0.15920	0.06206	-2.565	0.0132	*
d_6	0.11348	0.05511	2.059	0.0445	*

	Valore	Gdl
Devianza nulla	59.461	54
Devianza residua	48.311	52

Figura 2.19: Frequenza fumatori - regione Emilia Romagna; serie osservata VS serie stimata



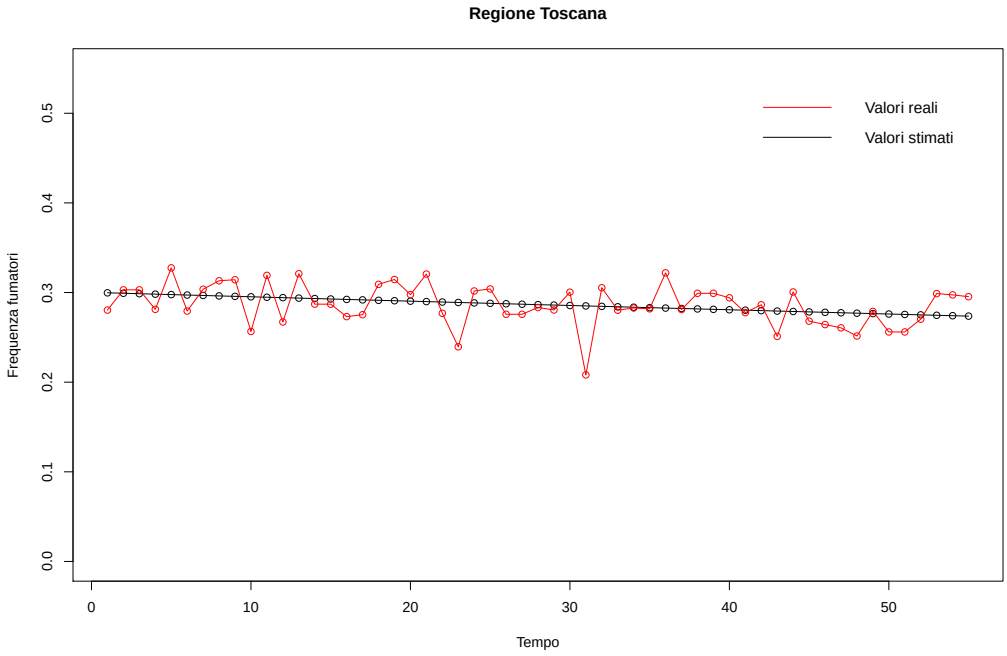
La frazione di fumatori riscontrata tra gli abitanti della Toscana presenta un lieve trend lineare decrescente, mentre in nessuna mensilità viene riscontrato qualche andamento significativamente particolare. La variabile inserita per modellare il trend riesce a spiegare circa il 12% della variabilità totale (tabella 2.24 e grafico 2.20).

Tabella 2.24: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Toscana. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.8464571	0.0275317	-30.745	< 2e-16	***
α_1	-0.0023548	0.0008829	-2.667	0.0101	*

	Valore	Gdl
Devianza nulla	41.259	54
Devianza residua	36.429	52

Figura 2.20: Frequenza fumatori - regione Toscana; serie osservata VS serie stimata



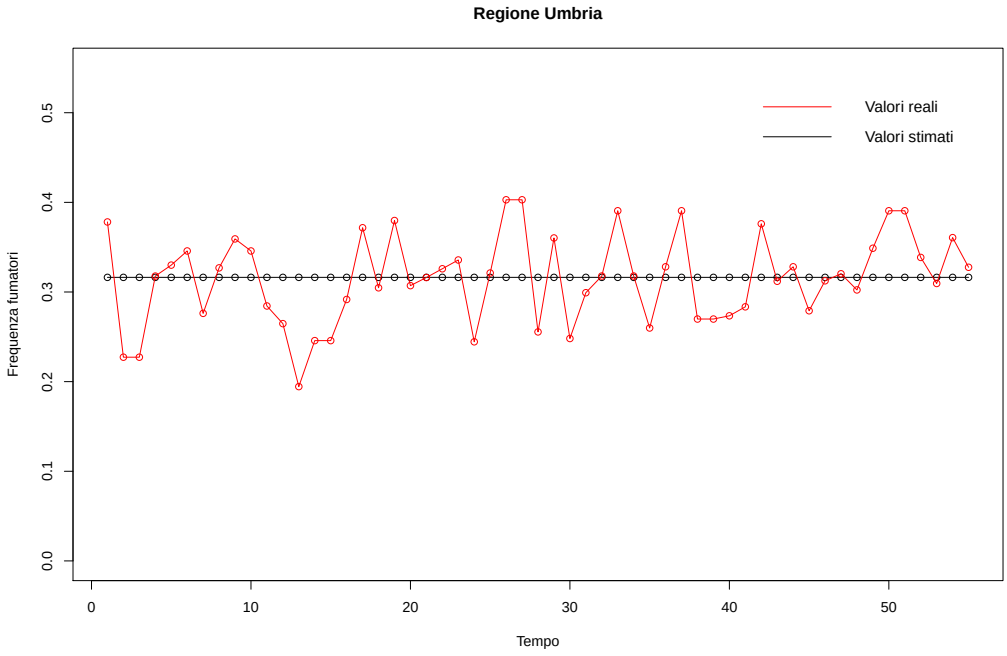
I dati rilevati nella regione Umbria presentano una marcata variabilità; in questo caso però, non può essere motivata dalla numerosità campionaria, poichè essa non risulta essere particolarmente bassa. Tale variabilità rientra completamente nella componente d'errore, visto che il modello più adeguato che abbiamo individuato presenta solamente l'intercetta (tabella 2.25 e grafico 2.21).

Tabella 2.25: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Umbria. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.77068	0.03105	-24.82	< 2e-16	***

	Valore	Gdl
Devianza nulla	77.485	54
Devianza residua	77.485	54

Figura 2.21: Frequenza fumatori - regione Umbria; serie osservata VS serie stimata



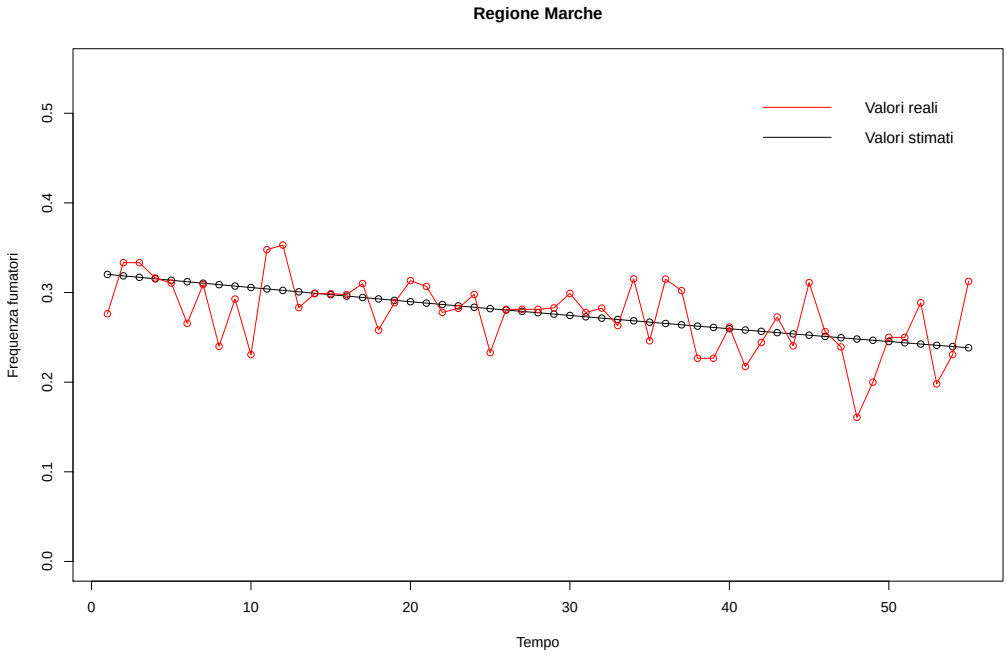
La serie che si riferisce ai fumatori marchigiani presenta un marcato trend lineare decrescente: la variabile inserita nel modello per rappresentarlo riesce a spiegare più del 36% di variabilità. Non vengono invece rilevate variabili *dummy* aventi parametri significativamente diversi da zero (tabella 2.26 e grafico 2.22).

Tabella 2.26: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Marche. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.745096	0.043146	-17.269	< 2e-16	***
α_1	-0.007572	0.001361	-5.563	8.91e-07	***

	Valore	Gdl
Devianza nulla	49.470	54
Devianza residua	31.327	53

Figura 2.22: Frequenza fumatori - regione Marche; serie osservata VS serie stimata



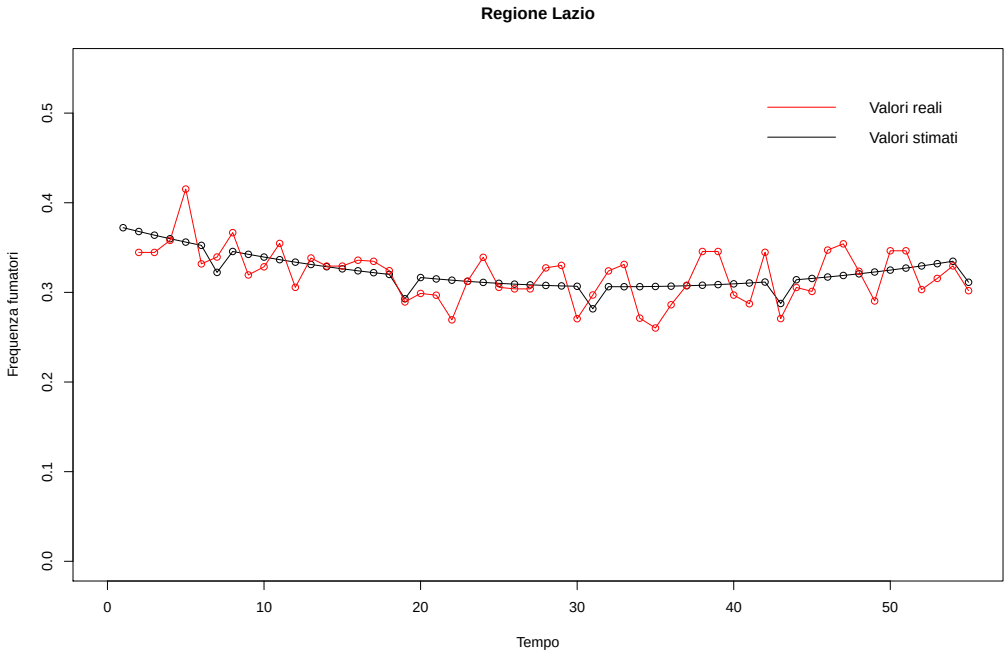
Il Lazio presenta una dato mancante, in corrispondenza di giugno 2007. Il modello stimato sulla base delle altre mensilità presenta una componente di trend quadratico che descrive una parabola con concavità verso l'alto, e una variabile *dummy* riferita al mese di dicembre, mese in cui la frequenza di fumatori risulta più bassa. I tre regressori inseriti spiegano circa il 33% di variabilità (tabella 2.27 e grafico 2.23).

Tabella 2.27: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Lazio. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.5041581	0.0554171	-9.098	3.51e-12	***
α_1	-0.0190895	0.0043031	-4.436	5.04e-05	***
α_2	0.0002908	0.0000721	4.034	0.000188	***
$d_1 2$	-0.1196695	0.0549157	-2.179	0.034059	*

	Valore	Gdl
Devianza nulla	54.437	53
Devianza residua	36.403	50

Figura 2.23: Frequenza fumatori - regione Lazio; serie osservata VS serie stimata



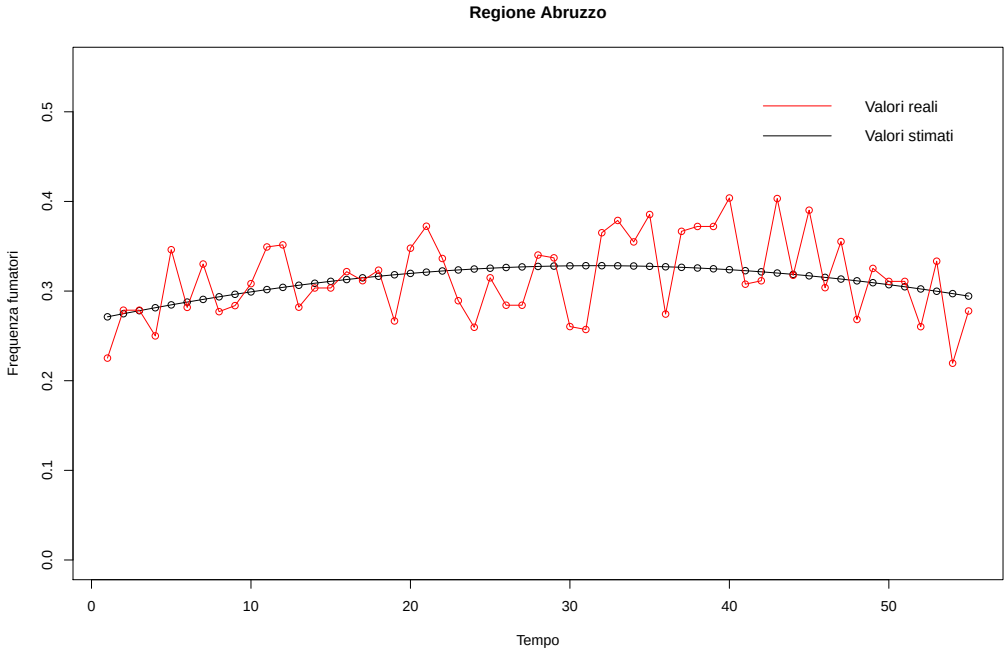
I dati che si riferiscono alla regione Abruzzo possono essere modellati utilizzando una componente di trend quadratico che descrive un andamento dapprima crescente e successivamente decrescente: la parabola ha dunque concavità verso il basso. Non si registrano invece ”effetti di mese” e la variabilità spiegata si aggira intorno al 15% (tabella 2.28 e grafico 2.24).

Tabella 2.28: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Abruzzo. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.0062725	0.0733698	-13.715	< 2e-16	***
α_1	0.0183664	0.0063659	2.885	0.00568	**
α_2	-0.0002903	0.0001128	-2.572	0.01300	*

	Valore	Gdl
Devianza nulla	43.402	54
Devianza residua	36.921	52

Figura 2.24: Frequenza fumatori - regione Abruzzo; serie osservata VS serie stimata



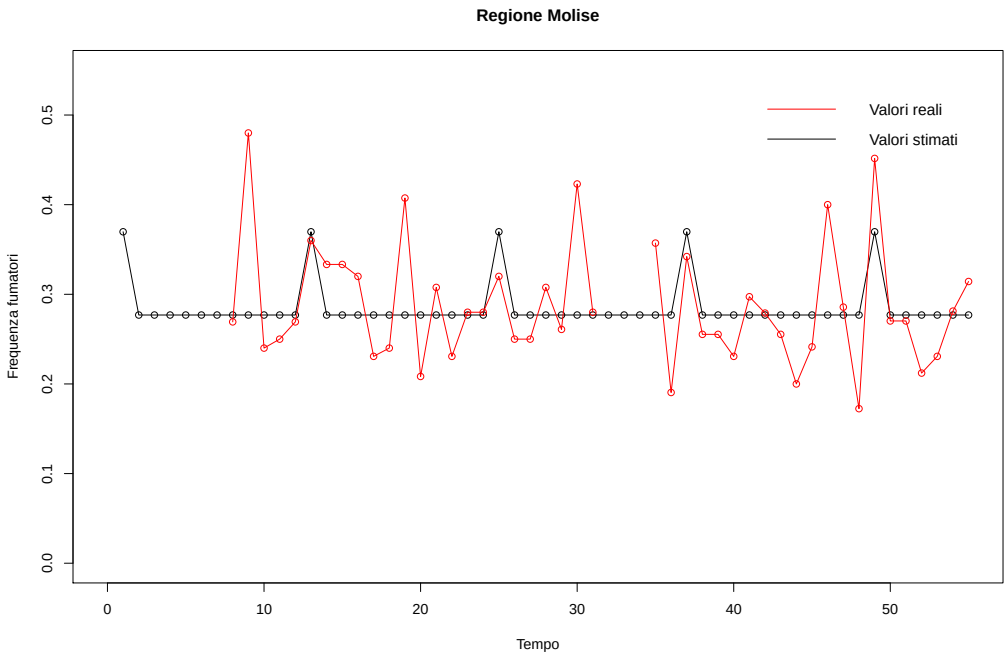
Il Molise presenta dieci mensilità con dati mancanti ed il modello stimato sulla base delle restanti presenta trend costante ed un "effetto di mese" relativo a giugno, mese nel quale si registra una frazione di fumatori superiore alla media. La sola variabile *dummy* presente spiega più del 15% di variabilità (tabella 2.29 e grafico 2.25).

Tabella 2.29: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Molise. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.95990	0.04761	-20.161	< 2e-16	***
d_6	0.42660	0.15151	2.816	0.00732	**

	Valore	Gdl
Devianza nulla	28.350	44
Devianza residua	23.956	43

Figura 2.25: Frequenza fumatori - regione Molise; serie osservata VS serie stimata



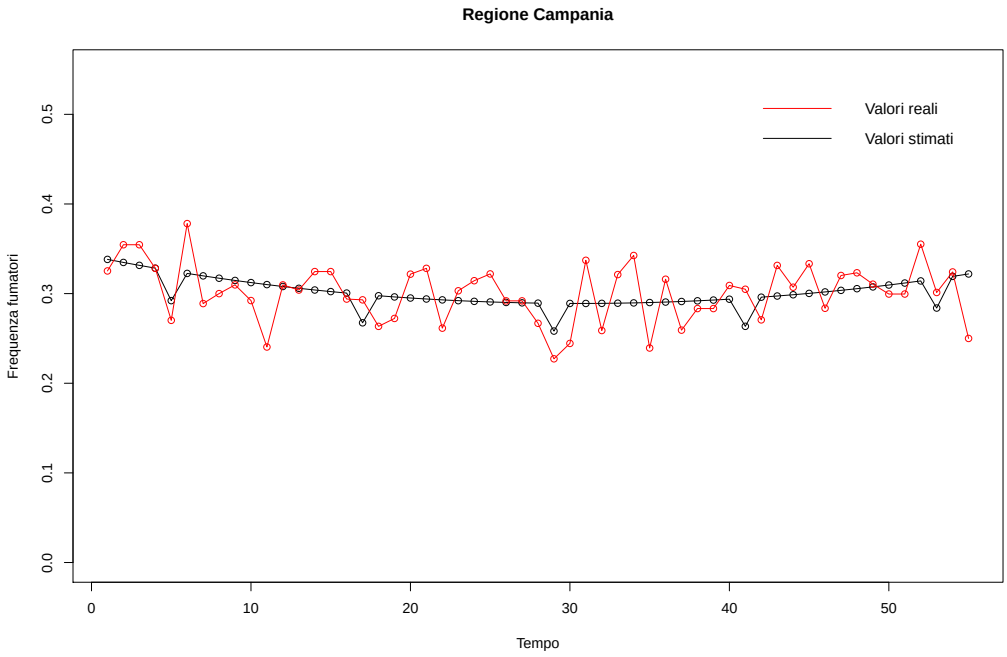
Per la Campania scegliamo una componente di trend quadratico, rappresentata da una parabola con concavità verso l'alto, e una variabile *dummy* riferita al mese di ottobre, in cui la frequenza di fumatori è inferiore rispetto a quanto descritto dalla componente di trend. Con questo modello riusciamo a spiegare poco più del 25% di variabilità (tabella 2.30 e grafico 2.26).

Tabella 2.30: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Campania. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.6556849	0.0533314	-12.295	< 2e-16	***
α_1	-0.0159737	0.0046484	-3.436	0.00118	**
α_2	0.0002609	0.0000836	3.121	0.00296	**
d_10	-0.1558121	0.0675206	-2.308	0.02511	*

	Valore	Gdl
Devianza nulla	75.157	54
Devianza residua	55.919	51

Figura 2.26: Frequenza fumatori - regione Campania; serie osservata VS serie stimata



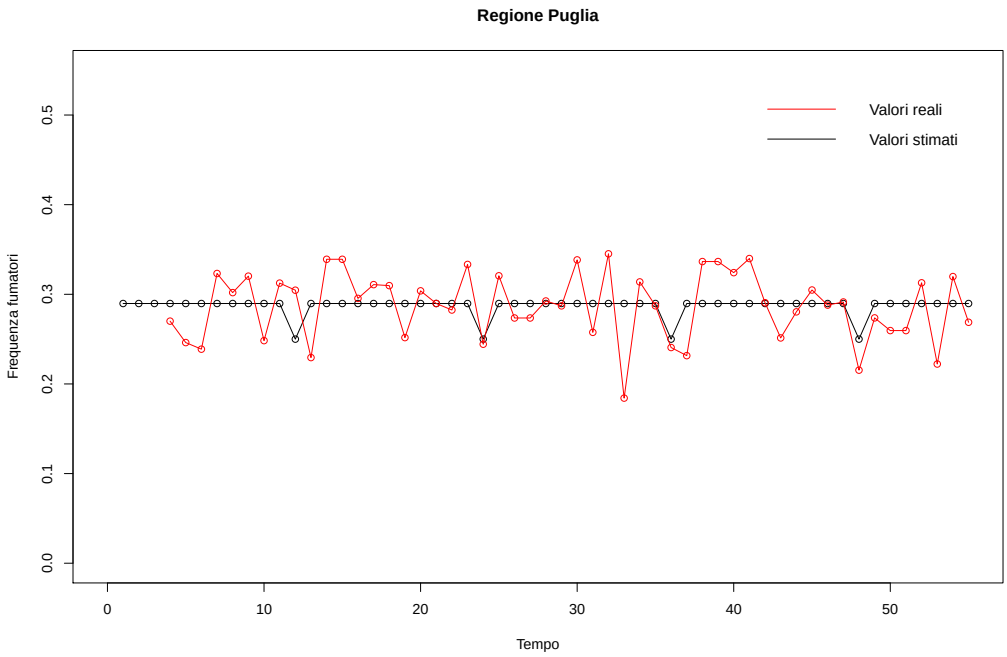
Per quanto riguarda la regione Puglia, mancano i valori dei mesi di giugno, luglio e agosto 2007. Il modello scelto per la serie prevede trend costante e l'inserimento della variabile *dummy* relativa a maggio, mese in cui la frazione di fumatori è più bassa rispetto alla media. La variabilità spiegata sfiora soltanto l'8% (tabella 2.31 e grafico 2.27).

Tabella 2.31: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Puglia. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.89667	0.02594	-34.561	< 2e-16	***
d_5	-0.20194	0.09763	-2.068	0.0438	*

	Valore	Gdl
Devianza nulla	67.191	51
Devianza residua	61.842	50

Figura 2.27: Frequenza fumatori - regione Puglia; serie osservata VS serie stimata



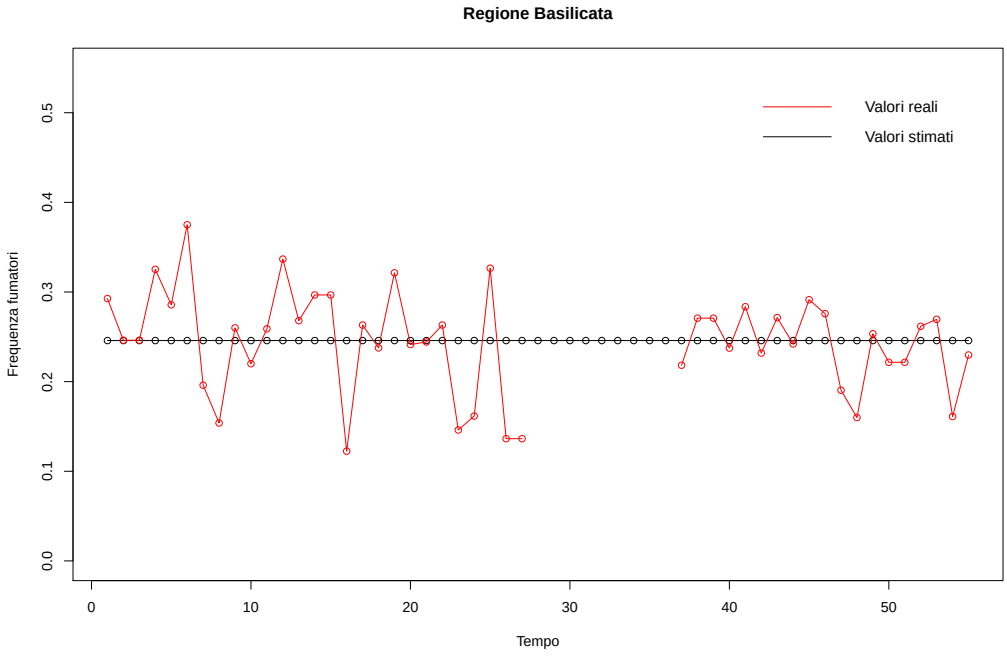
Sono nove le mensilità mancanti nei dati relativi alla regione Basilicata e nel modellare la serie costruita con i valori dei mesi restanti non riusciamo ad individuare alcun parametro significativamente diverso da zero, nè relativo alla componente di trend, nè relativo alle variabili *dummy* (tabella 2.32 e grafico 2.28).

Tabella 2.32: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Basilicata. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.12100	0.03941	-28.45	< 2e-16	***

	Valore	Gdl
Devianza nulla	70.167	45
Devianza residua	70.167	45

Figura 2.28: Frequenza fumatori - regione Basilicata; serie osservata VS serie stimata



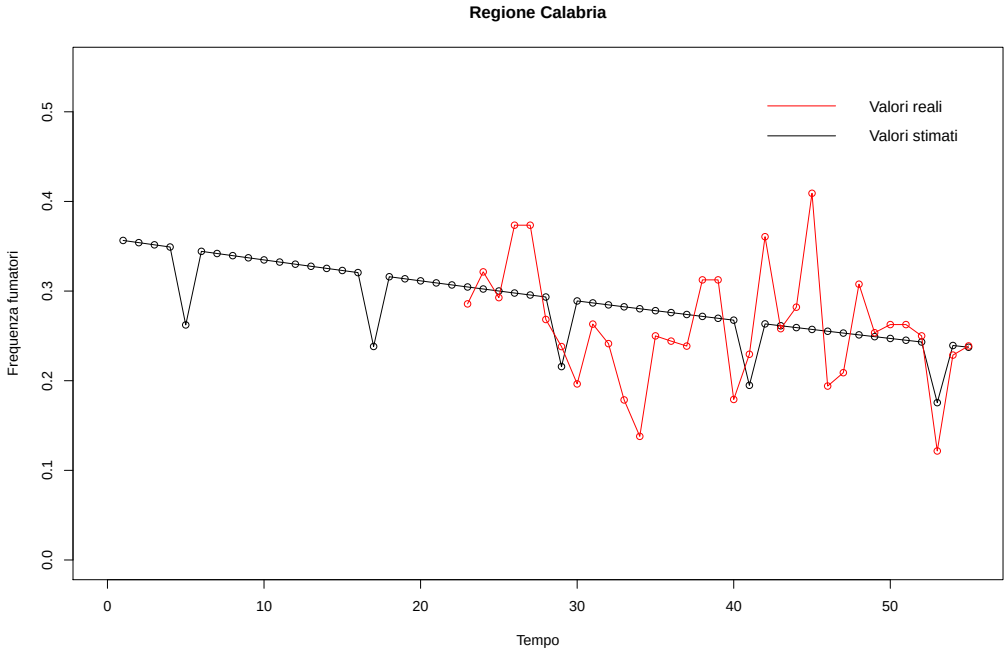
Le rilevazioni riguardanti la Calabria iniziano dal mese di aprile 2009; mancano dunque i dati di ben 22 mensilità. La serie presenta un trend lineare decrescente e nei mesi di ottobre la frazione di fumatori pare essere significativamente più bassa. Le variabili inserite come regressori spiegano quasi il 25% di variabilità (tabella 2.33 e grafico 2.29).

Tabella 2.33: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Calabria. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.580022	0.204444	-2.837	0.00808	**
α_1	-0.010677	0.005106	-2.091	0.04511	*
d_10	-0.400930	0.182125	-2.201	0.03554	*

	Valore	Gdl
Devianza nulla	42.166	32
Devianza residua	31.774	30

Figura 2.29: Frequenza fumatori - regione Calabria; serie osservata VS serie stimata



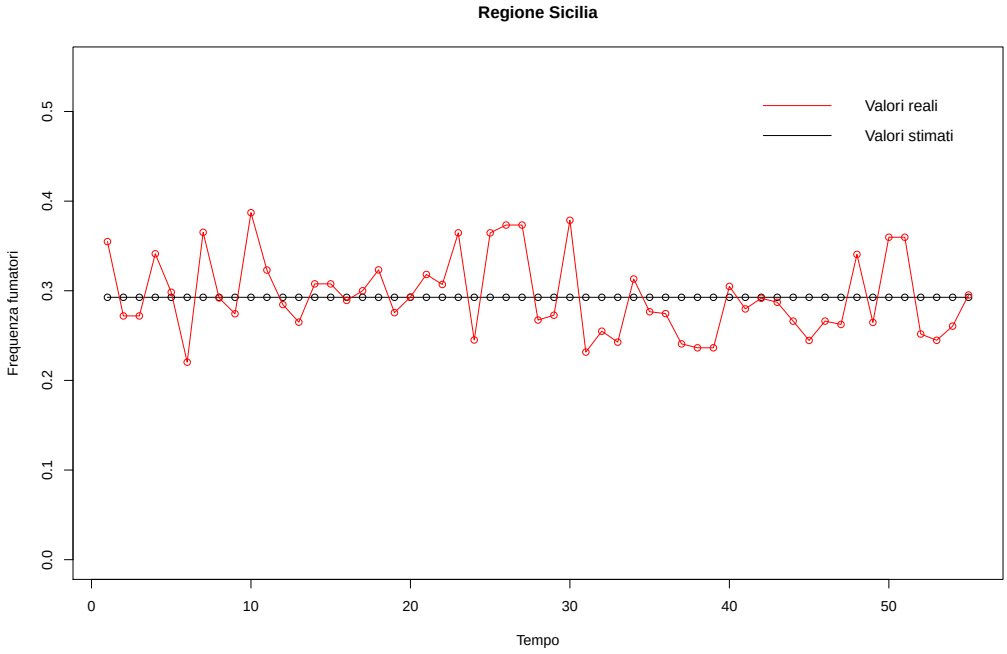
Il modello adottato per i dati relativi alla Sicilia presenta solo l'intercetta e tutta la variabilità è lasciata come componente d'errore. Il trend stimato è dunque costante e non si rilevano particolari "effetti di mese" (tabella 2.34 e grafico 2.30).

Tabella 2.34: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Sicilia. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-0.88227	0.02775	-31.79	<2e-16	***

	Valore	Gdl
Devianza nulla	58.001	54
Devianza residua	58.001	54

Figura 2.30: Frequenza fumatori - regione Sicilia; serie osservata VS serie stimata



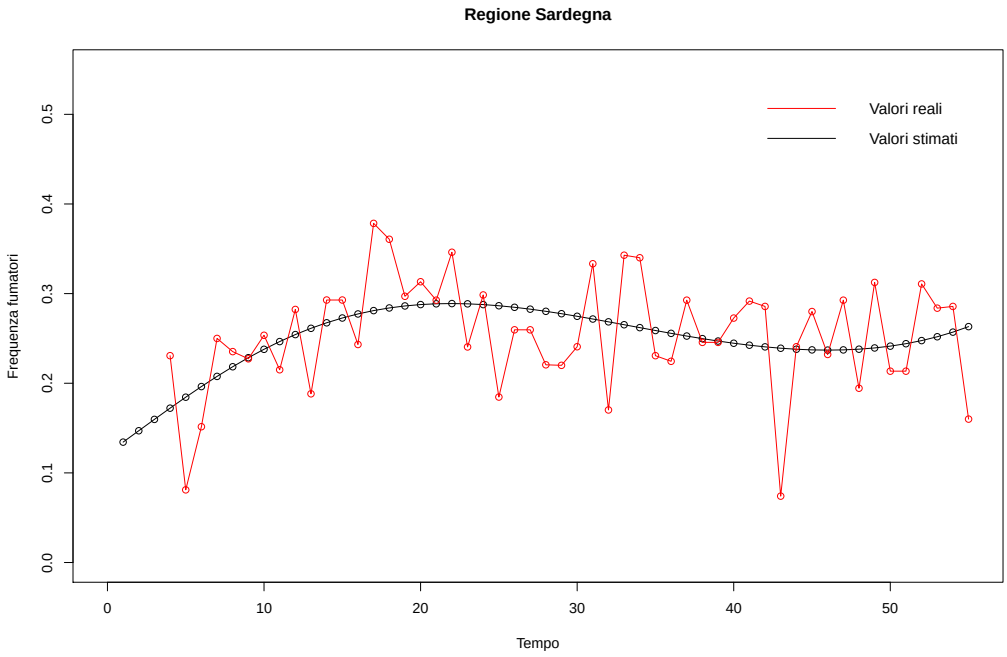
Mancano i dati dei mesi di giugno, luglio e agosto 2007 per quanto riguarda la Sardegna. Il modello proposto presenta una componente di trend polinomiale di terzo grado e nessuna variabile *dummy*; la variabilità spiegata è del circa il 16% (tabella 2.35 e grafico 2.31).

Tabella 2.35: Stime dei parametri e devianza del modello di regressione Beta-Binomiale (trend + stagionalità con variabili *dummy*) riferito alla frazione di fumatori della regione Sardegna. Approccio *forward*

	Stima	Std. Error	t value	*p-value*	Significatività
α_0	-1.976e+00	3.365e-01	-5.872	3.92e-07	***
α_1	1.165e-01	4.148e-02	2.809	0.00716	***
α_2	-3.922e-03	1.509e-03	-2.598	0.01241	*
α_3	3.847e-05	1.640e-05	2.346	0.02316	*

	Valore	Gdl
Devianza nulla	56.525	51
Devianza residua	47.142	48

Figura 2.31: Frequenza fumatori - regione Sardegna; serie osservata VS serie stimata



2.4 Considerazioni

Riportiamo alcune considerazioni dopo l'analisi dei risultati ottenuti seguendo l'approccio dei modelli di regressione lineari generalizzati (ed in particolare del

GLM basato sulla distribuzione Beta-Binomiale) applicati singolarmente ad ogni strato di nostro interesse e costruiti sulla base di regressori temporali.

- Le strutture cambiano molto da strato a strato, a causa della numerosità campionaria variabile e dell'irregolarità del fenomeno nel tempo; a fronte di quest'ultimo aspetto, in tutti i casi esposti la variabilità spiegata dal modello non raggiunge il 50%, a volte anzi è nulla, poichè il modello presenta solamente l'intercetta come parametro significativamente diverso da zero.

- Sembra plausibile, nella maggior parte dei casi, la presenza di una tendenza di fondo, per lo più di tipo lineare o quadratica.

- Non riscontriamo la presenza di una componente stagionale periodica regolare ed anche gli "effetti di mese" sono rari e, se presenti, risultano diversi da caso a caso (mesi diversi, oppure parametri con segni opposti per lo stesso mese): le oscillazioni rilevate nelle serie storiche considerate non sono dunque riconducibili a comportamenti che si ripetono ciclicamente.

Capitolo 3

MODELLAZIONE MULTILIVELLO

3.1 Modelli multilivello: a cosa servono e come sono formulati

Spesso ci troviamo ad analizzare dati che presentano una struttura gerarchica o clusterizzata [Goldstein 1999]. Nel caso preso in esame, notiamo che gli intervistati possono essere raggruppati secondo determinati criteri quali classe d'età, sesso, livello d'istruzione, regione d'appartenenza; dunque, possiamo costruire diverse partizioni della popolazione di riferimento, in base ad uno dei criteri citati, o a combinazioni di essi.

In precedenza, abbiamo modellato la variabile d'interesse (frazione di fumatori) in funzione del tempo, in due direzioni:

1. riferendoci al campione totale, senza suddividere in categorie gli intervistati e di conseguenza ignorando eventuali differenze tra gruppi di individui (*complete-pooling*);

2. costruendo un modello di regressione per ogni gruppo di riferimento (*no-pooling*; per esempio, tre modelli differenti, per ognuna delle classi d'età individuate); questo approccio, all'estremo opposto rispetto al primo, tende ad ingigantire le differenze tra i gruppi e a far apparire quindi i vari andamenti individuati più diversi di quanto non lo siano realmente.

La regressione multilivello (*partial-pooling*) può essere pensata come un metodo per raggiungere un compromesso tra i due estremi appena descritti e sviluppati nel capitolo precedente [Gelman, Hill 2006]. In particolare, i parametri contenuti in un modello multilivello (intercetta e/o coefficienti dei regressori) possono variare da gruppo a gruppo secondo una determinata distribuzione di probabilità.

3.1.1 Struttura dei modelli lineari multilivello

3.1.1.1 Modelli di base

Concentriamoci sulla struttura prevista per costruire un modello lineare multilivello; iniziamo con il presentare un modello semplice e proseguiamo con una formulazione più complessa che ci sarà utile per giungere al modello più adeguato per il caso preso in esame.

Supponiamo di avere n osservazioni (identificate dall'indice $i = 1, \ldots, n$), che possiamo riunire in J gruppi (identificati dall'indice $j = 1, \ldots, J$), e di utilizzare come unico regressore la variabile X, definita a livello individuale. La struttura presa in considerazione presenta dunque due livelli ed una variabile di regressione definita a livello 1, ossia a livello di unità statistica. Se ipotizziamo che ogni gruppo si differenzi dagli altri per l'intercetta e non per il coefficiente angolare della retta di regressione (e quindi che il parametro associato al regressore non cambi da gruppo a gruppo, mantenendo invariata la relazione tra esso e la variabile di interesse), il modello può essere formulato come

$$y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i$$

dove

- l'indice $j[i]$ codifica il gruppo di appartenenza: l'i-esima unità appartiene al j-esimo gruppo;
- $y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$;

- $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$, o, in alternativa, $\alpha_j = \mu_\alpha + \eta_j$, con $\eta_j \sim N(0, \sigma_\alpha^2)$; quest'ultima formulazione è utile per comprendere meglio le differenze tra i gruppi;
- σ_y^2 è la varianza all'interno di un gruppo, che per semplicità assumiamo costante;
- σ_α^2 è la varianza tra le medie campionarie di ogni gruppo;
- se $\sigma_\alpha \to \infty$ ci si riconduce alla metodologia *no-pooling*;
- se $\sigma_\alpha \to 0$ ci si riconduce al *complete-pooling*;
- una stima approssimata di α_j può essere espressa come una media pesata della stima *no-pooling* del j-esimo gruppo $(\bar{y}_j - \beta\bar{x}_j)$ e la media μ_α:

$$\hat{\alpha_j} \approx \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}(\bar{y}_j - \beta\bar{x}_j) + \frac{\frac{1}{\sigma_\alpha^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}\mu_\alpha;$$

  si può notare che, se $n_j \to 0$ la stima del parametro α_j è prossima a μ_α, mentre se $n_j \to \infty$ la stima è prossima alla media campionaria del j-esimo gruppo.

Se supponiamo che anche il coefficiente angolare sia variabile da gruppo a gruppo, il modello assume la forma:

$$y_i \sim N(\alpha_{j[i]} + \beta_{j[i]}x_i, \sigma_y^2), \qquad i = 1, \ldots, n$$

$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\alpha \\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right), \qquad j = 1, \ldots, J$$

dove ρ è una misura della correlazione tra intercetta e coefficiente angolare.

La formulazione appena presentata ci sarebbe utile per modellare una componente di trend lineare (l'unico regressore necessario è un vettore per indicare i periodi di tempo), considerando un solo criterio di stratificazione. Ma potremmo aver bisogno di un modello più complesso, in due direzioni:

1. potremmo avere la necessità di inserire più di un predittore, ad esempio per modellare una componente di trend polinomiale di secondo o terzo grado;

2. potremmo voler combinare più criteri di stratificazione per formare una partizione (per esempio, categorizzare gli individui in base alla classe d'età *e* al sesso), giungendo dunque ad un modello non nidificato, poichè i livelli di raggruppamento non formano una gerarchia.

3.1.1.2 Modelli con due o più regressori

Ipotizziamo di inserire nel modello due regressori e di considerare variabili da gruppo a gruppo sia l'intercetta sia i coefficienti dei due predittori; il modello può essere scritto nella forma standard

$$y_i \sim N(\beta_{0j[i]} + \beta_{1j[i]}x_{1i} + \beta_{2j[i]}x_{2i}, \sigma_y^2)$$

oppure tramite la notazione matriciale

$$y_i \sim N(X_i B_{j[i]}, \sigma_y^2), \qquad i = 1, \ldots, n$$

$$\beta_j \sim N(M_B, \Sigma_B), \qquad j = 1, \ldots, J$$

dove

- X è la matrice $n \times 3$ dei predittori: la prima colonna di X è un vettore di tutti 1 mentre le altre due colonne contengono i valori assunti rispettivamente dalla variabile X_1 e X_2. X_i è il vettore di lunghezza 3 che rappresenta l'i-esima riga di X e $X_i B_{j[i]}$ è $\beta_{0j[i]} + \beta_{1j[i]}x_{1i} + \beta_{2j[i]}x_{2i}$.

- B è la matrice $3 \times J$ dei coefficienti di regressione; B_j è il vettore di lunghezza 3 che rappresenta la j-esima colonna di B.

- $M_B = (\mu_{\beta_0}, \mu_{\beta_1}, \mu_{\beta_2})$ è il vettore di lunghezza 3 che rappresenta le medie delle distribuzioni dei tre parametri considerati.

- Σ_B è la matrice 3×3 di varianze e covarianze dei parametri.

Quando all'interno del modello il numero K dei coefficienti variabili è maggiore di 2, la procedura di stima è più complessa, soprattutto per quanto riguarda il parametro di correlazione ρ. Per stimare la matrice Σ_B ci affidiamo alla distribuzione Inverse-Wishart, conveniente dal punto di vista computazionale ma difficile da interpretare [Gelman, Hill 2006]. Accenniamo solamente al fatto che la matrice di varianze e covarianze può essere espressa come

$$\Sigma_B = Diag(\xi)\mathcal{Q}Diag(\xi),$$

dove ξ_k è un vettore di parametri di scala e $(Q) \sim Inv - Wishart_{K+1}(I)$; questa formulazione ha il pregio di rispettare il vincoli che impongono che Σ_B sia definita positiva e che ρ assuma valori all'interno dell'intervallo $[-1; 1]$.

3.1.1.3 Modelli non annidati

Supponiamo di voler considerare due criteri di stratificazione contemporaneamente, che non formano una gerarchia (come invece succederebbe se ad esempio ci riferissimo a comune, provincia, regione), e un unico regressore definito a livello di unità statistica. Il modello può essere espresso come

$$y_i \sim N(\alpha_{j[i],k[i]} + \beta_{j[i]k[i]}x_i, \sigma_y^2), \qquad i = 1, \ldots, n.$$

Le formulazioni e le spiegazioni fornite precedentemente per i parametri restano valide anche in questo caso.

3.1.2 Struttura dei modelli logistici multilivello

Poichè ciò che ci interessa modellare è la frazione di fumatori, ricerchiamo anche all'interno della regressione multilivello un modello che ci permetta di raggiungere quest'obiettivo. Ci affidiamo alla regressione logistica, che adotta la distribuzione Binomiale per rappresentare la variabile d'interesse e la funzione logistica per mettere in relazione Y con il predittore lineare considerato.

In precedenza ci eravamo affidati, per una stima più precisa dei dati, alla distribuzione Beta-Binomiale, la cui implementazione in R prevedeva la stima del parametro di dispersione; la funzione `lmer` usata in R per stimare i modelli multilivello, non prevede però la possibilità di affidarsi a questa distribuzione e dunque il parametro di dispersione sarà costante e pari a 1.

Se consideriamo un unico regressore ed un unico livello di stratificazione, possiamo stimare la probabilità di successo come segue:

$$y_i = logit^{-1}(\alpha_{j[i]} + \beta_{j[i]}x_i) = \frac{e^{\alpha_{j[i]}+\beta_{j[i]}x_i}}{1 + e^{\alpha_{j[i]}+\beta_{j[i]}x_i}},$$

con, come prima,

$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\alpha \\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right), \qquad j = 1, \ldots, J.$$

3.2 Metodi di valutazione dei modelli

Per valutare la bontà di un modello ci affidiamo all'analisi dei residui, attraverso test sulla normalità e sulle autocorrelazioni, essendo in presenza di serie storiche. Per confrontare tra di loro i modelli ci affidiamo invece ai criteri AIC e BIC.

3.2.1 Analisi dei residui

L'analisi dei residui stimati permette di verificare l'adeguatezza del modello scelto per stimare le componenti deterministiche e per svolgere delle previsioni. Il modello risulta adeguato se l'ipotesi di casualità dei residui è confermata, cioè se la serie dei residui è stata ragionevolmente generata da un processo *white noise* secondo cui il residuo al tempo t, r_t, ha queste proprietà:

$$E(r_t) = 0, \qquad var(r_t) = \sigma_r^2, \qquad cov(r_s, r_t) = 0, \qquad \forall t, t \neq s.$$

Alcuni test si basano sull'ulteriore ipotesi di normalità dei residui, altri invece non pongono restrizioni sulla forma della distribuzione dei r_t; questi ultimi, di

cui il test sulle autocorrelazioni fa parte, godono di una maggiore generalità ma solitamente sono meno potenti, ossia presentano una più alta probabilità di non rifiuto di H_0 nel caso in cui essa non sia vera [Di Fonzo, Lisi 2005].

3.2.1.1 Test di normalità

Per valutare la plausibilità dell'ipotesi di normalità dei residui, dopo un'opportuna standardizzazione di essi, ci affideremo ad un'analisi grafica ed al test di Shapiro-Wilk, basato sul confronto di sue stimatori alternativi della varianza: uno stimatore non parametrico ed uno parametrico (varianza campionaria).

3.2.1.2 Test sulle autocorrelazioni

I valori su cui puntiamo la nostra attenzione rappresentano delle proporzioni calcolate in riferimento a determinati periodi di tempo. Possiamo dunque pensare questi dati in qualità di realizzazioni di un processo stocastico ossia di una collezione di variabili casuali indicizzate dal tempo:

$$\{Y_t, t = t_1, t_2, \dots\}.$$

Possiamo definire le funzioni di media, varianza e autocovarianza, che al variare di t assumono valori differenti:

$$\mu_t = E(Y_t), \qquad \sigma_t^2 = E(Y_t - \mu_t)^2, \qquad \gamma_{t_1,t_2} = E[(Y_{t1} - \mu_{t1})(Y_{t2} - \mu_{t2})].$$

La funzione γ_{t_1,t_2} rappresenta la covarianza tra variabili casuali del medesimo processo stocastico spaziate tra di loro da uno sfasamento temporale pari a $k = |t_2 - t_1|$. Se effettuiamo una normalizzazione, arriviamo a definire la funzione di autocorrelazione (ACF, acronimo di *AutoCorrelation Function*):

$$\rho_{t_1,t_2} = \frac{\gamma_{t_1,t_2}}{\sigma_{t_1}\sigma_{t_2}}$$

É stato dimostrato che i coefficienti di autocorrelazione campionari (indicati, più sinteticamente, con ρ_k) di una successione di n valori generati da un processo

white noise, per n sufficientemente grande, si distribuiscono approssimativamente come una Normale di media nulla, varianza pari a circa $\frac{1}{n}$ e non sono tra loro correlati [Di Fonzo, Lisi 2005].

Si respinge l'ipotesi di una serie generata da un *white noise* quando si riscontrano valori dei coefficienti di autocorrelazione stimati esterni all'intervallo $[-\frac{z_{1-\alpha/2}}{\sqrt{n}}, +\frac{z_{1-\alpha/2}}{\sqrt{n}}]$, dove $z_{1-\alpha/2}$ è il percentile di livello $1-\alpha/2$ di una $N(0,1)$ e α è il livello di significatività prescelto, solitamente pari a 0.05.

3.2.2 Criteri per la selezione dei modelli

Se vogliamo operare una scelta tra due o più modelli, confrontiamo i valori che assumono due misure relative alla bontà di stima: l'AIC (*Akaike Information Criterion*) ed il BIC (*Bayesian Information Criterion*). Entrambi questi criteri esprimono un trade-off tra la verosimiglianza ed il numero di parametri: l'aggiunta di parametri provoca un aumento del valore assunto dalla verosimiglianza, il che si traduce in una miglior rappresentazione dei dati da parte del modello; d'altro canto, aumenta la complessità e si rischia di incorrere nell'*overfitting*.

3.2.2.1 AIC: *Akaike Information Criterion*

L'AIC è un test statistico basato sul concetto di entropia e offre una misura relativa della quantità di informazione persa quando un modello è usato per descrivere la realtà. Matematicamente, il test di verifica delle informazioni di Akaike è così definito:

$$AIC = -2 \ln L + 2k$$

dove:

- L è il valore della massima verosimiglianza del modello.
- k è il numero di parametri inseriti nel modello;

Nella scelta tra due o più modelli, si preferisce quello che presenta l'AIC più basso.

3.2.2.2 BIC: *Bayesian Information Criterion*

Il BIC risulta molto simile all'AIC, come si nota dalla sua formulazione, ma penalizza maggiormente l'aggiunta di parametri nel modello.

$$BIC = -2 \ln L + k \ln n$$

dove:

- L è il valore della massima verosimiglianza del modello.
- k è il numero di parametri inseriti nel modello;
- n è il numero di osservazioni;

Anche in questo caso, si preferisce il modello con il BIC più basso.

Nelle nostre analisi, ci affideremo soprattutto a quest'ultima misura, che penalizza in modo più marcato l'aggiunta di parametri, poichè a fronte di un n (nel nostro caso, periodi di tempo) non elevato, è preferibile optare per un modello con un numero di parametri non eccessivamente alto.

3.3 Applicazione ai dati relativi al fumo

Per analizzare l'andamento temporale della percentuale di fumatori all'interno di determinate sottopopolazioni costruiremo dei modelli multilivello applicati alla regressione logistica. Elenchiamo gli elementi che costituiranno i modelli, caratterizzati da parametri associati ai regressori ed intercetta variabili [1]:

[1] Poichè stiamo trattando modelli logistici, le stime dei parametri *intercetta* e *coefficiente angolare*, termini usati in seguito per comodità, non si riferiscono esattamente ai valori assunti realmente dalle rette risultanti dal modello, bensì sono una funzione di essi e rappresentano le stime dei parametri del predittore lineare.

- regressori temporali per indicare i periodi di rilevazione, ovvero: $t = 1, \ldots, n$ per modellare una componente di trend lineare; per modellare invece un trend quadratico, aggiungiamo un secondo regressore, $t^2 = 1^2, \ldots, n^2$; e così via, ma noi ci fermeremo ad analizzare trend polinomiali di primo, secondo e terzo grado. Non inseriremo invece predittori per modellare la stagionalità, perchè dopo le analisi precedenti ci sentiamo di escludere la presenza di componenti stagionali significative;

- quattro variabili di stratificazione: classe d'età, sesso, livello d'istruzione, regione. Inizieremo col considerare un criterio di partizione alla volta; successivamente passeremo a modelli non annidati, in cui le sottopopolazioni verranno individuate da due variabili di stratificazione.

Nel caso di un modello con componente di trend lineare valuteremo, oltre che le differenze tra gruppi di una stessa partizione, anche le differenze tra le rette stimate per ciascuna sottopopolazione e la retta di regressione media determinata da μ_α e μ_β.

3.3.1 Modelli con un'unica variabile di stratificazione

Sviluppiamo dei modelli che, se ipotizziamo di rappresentare una tendenza di fondo tramite un polinomio di primo grado, assumono questa formulazione:

$$y_i = logit^{-1}(\alpha_{j[i]} + \beta_{j[i]} t_i) = \frac{e^{\alpha_{j[i]} + \beta_{j[i]} t_i}}{1 + e^{\alpha_{j[i]} + \beta_{j[i]} t_i}},$$

con

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right), \qquad j = 1, \ldots, J.$$

Oppure, possiamo esprimere i due parametri α_j e β_j nel seguente modo, per capire con maggiore chiarezza le differenze tra gruppo e gruppo:

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} = \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix} + \begin{pmatrix} \eta_j \\ \zeta_j \end{pmatrix}$$

con

$$\begin{pmatrix}\eta_j\\ \zeta_j\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_\alpha & \rho\sigma_\alpha\sigma_\beta\\ \rho\sigma_\alpha\sigma_\beta & \sigma^2_\beta\end{pmatrix}\right), \qquad j = 1, \dots, J.$$

3.3.1.1 Variabile di stratificazione: classe d'età

Partizioniamo gli intervistati secondo la classe d'età a cui appartengono, così da ottenere tre gruppi: 18-34 anni, 35-49 anni, 50-69 anni. Di media, i campioni rilevati ogni mese sono composti dal 28.33% di individui appartenenti alla prima fascia d'età, dal 34.55% di persone appartenenti alla seconda e dal 37.12% di persone con età compresa tra i 50 e i 69 anni.

Costruiamo un modello contenente una componente di trend lineare. La tabella 3.1 contiene i valori stimati di μ_α e μ_β, ossia delle medie dei parametri variabili, di σ^2_α, σ^2_β, $\rho\sigma_\alpha\sigma_\beta$ ed infine ρ, che rappresenta la correlazione tra l'intercetta ed il coefficiente angolare. Mediamente il trend è decrescente, e la correlazione pari a -1 sta a significare che ad un aumento del valore dell'intercetta (aumento del livello di "partenza" della percentuale di fumatori), corrisponde una proporzionale diminuzione del valore del coefficiente angolare: più la percentuale di fumatori iniziale è alta, più la decrescita, in questo caso, nel corso del tempo risulta marcata. Viceversa, ad una diminuzione del valore dell'intercetta corrisponde un proporzionale aumento del coefficiente angolare, con conseguente diminuzione della velocità di decrescita o eventuale cambiamento di segno del coefficiente, così da determinare un andamento crescente.

Tabella 3.1: Stime dei parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età

	Stima	Std. Error	t value	*p-value*	Significatività
μ_α	-0.8785529	0.1544723	-5.687	1.29e-08	***
μ_β	-0.0015318	0.0005963	-2.569	0.0102	*

σ^2_α	σ^2_β	$\rho\sigma_\alpha\sigma_\beta$	ρ
7.1253e-02	7.4353e-07	-2.301703e-04	-1

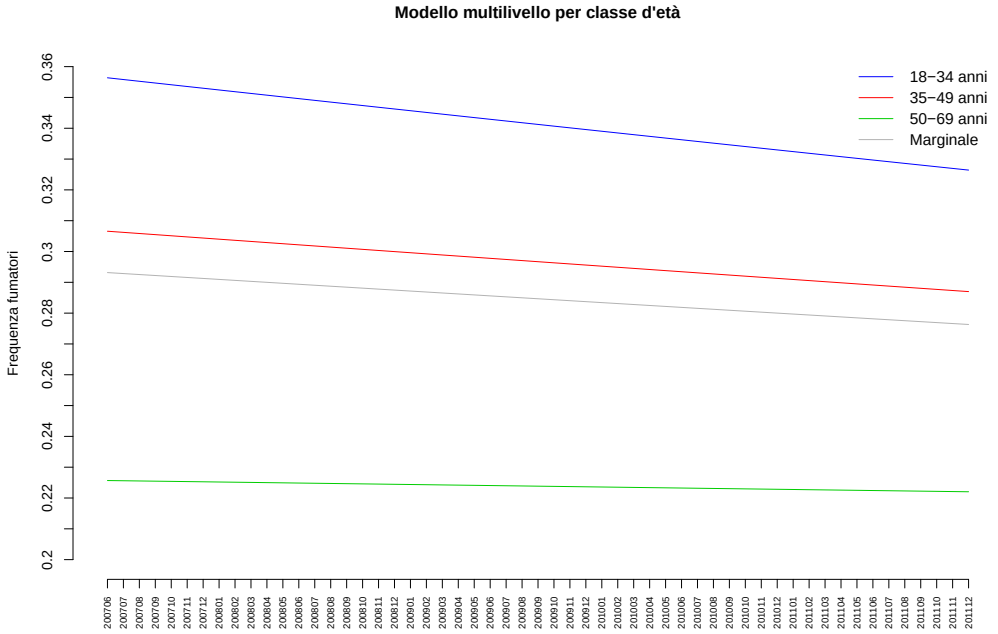
Nella tabella 3.2 sono invece contenuti gli scostamenti dai valori medi dei parametri, relativi ad ogni singolo gruppo ed il grafico 3.1 illustra il modello proposto, permettendoci di confrontare tra loro le rette di regressione stimate per ogni gruppo e di valutarle anche in relazione alla retta media stimata.

L'intercetta stimata per il gruppo 18-34 anni è decisamente più elevata rispetto alla media; in compenso, come spiegato prima, la decrescita è più evidente, infatti nel corso di tre anni e mezzo si è passati da una percentuale stimata del 35.64% al 32.64%. Tra gli individui di età compresa tra i 35 ed i 49 anni il livello di fumatori è lievemente al di sopra della media, mentre il gruppo 50-69 presenta una frazione di fumatori molto più bassa rispetto alla media e segue un andamento pressochè costante nel corso del tempo.

Tabella 3.2: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età

Gruppo	η_j	ζ_j
18-34	0.28989896	-0.0009364696
35-49	0.06414821	-0.0002072200
50-69	-0.35398668	0.0011434942

Figura 3.1: Rappresentazione del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età



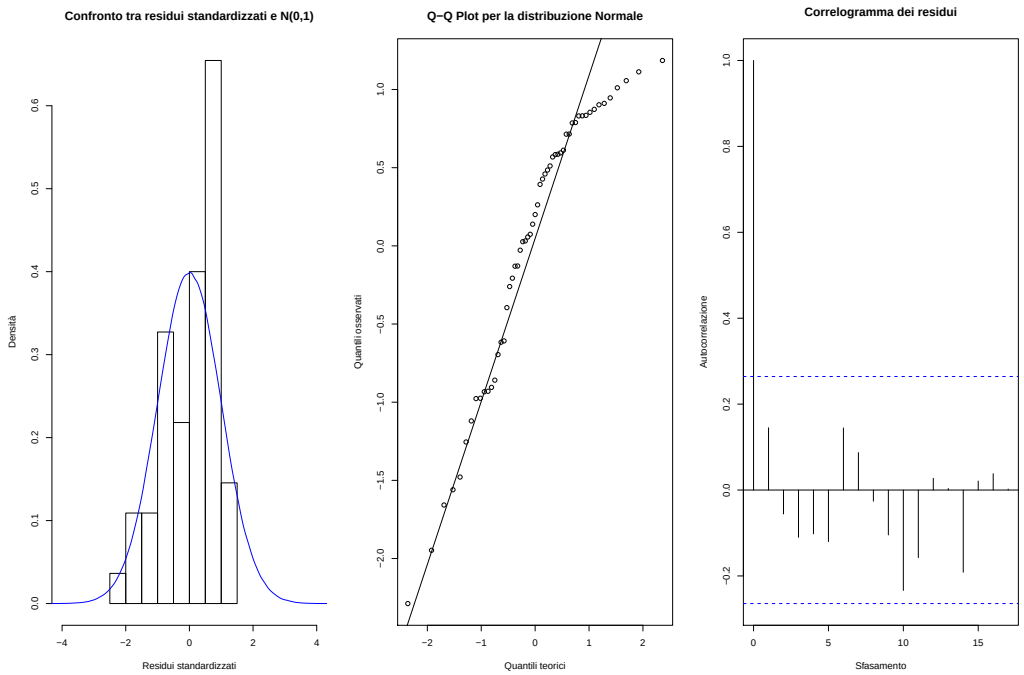
Dopo aver standardizzato i residui, effettuiamo l'analisi di essi all'interno di ogni gruppo, separatamente.

1. Classe d'età 18-34 anni: il test di Shapiro-Wilk presenta un *p-value* pari a 0.001349 e ci suggerisce dunque di rifiutare l'ipotesi di normalità. Dal grafico 3.2 possiamo riscontrare un certa irregolarità nella distribuzione dei residui ed anche la tabella 3.3 lo conferma: si nota un'asimmetria verso destra, confermato dal valore con segno positivo che assume la mediana, il che significa che più della metà (precisamente, il 60%) dei valori assunti dalla variabile di interesse nel modello sono sottostimati. Il grafico delle autocorrelazioni non presenta coefficienti significativamente diversi da zero, di conseguenza l'ipotesi di aleatorietà dei residui non viene rifiutata.

Tabella 3.3: *Summary* dei residui standardizzati - classe d'età 18-34 anni

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.288000	-0.656000	0.200200	0.007574	0.750800	1.186000

Figura 3.2: Analisi dei residui: normalità e autocorrelazioni - classe d'età 18-34 anni



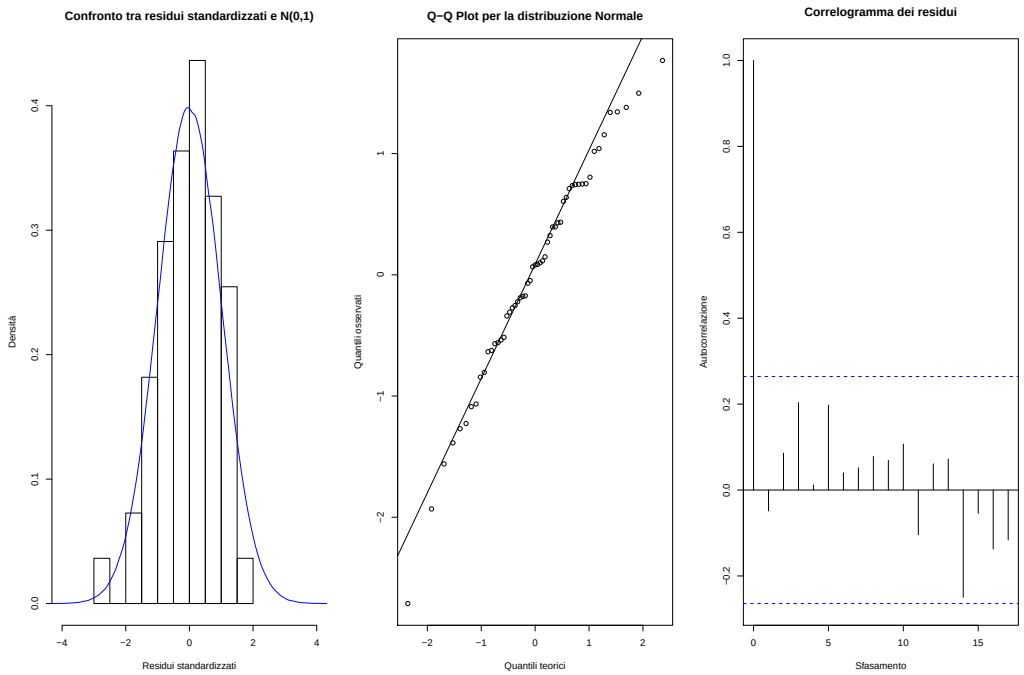
2. Classe d'età 35-49 anni: dal test di Shapiro-Wilk (*p-value* pari allo 0.5437) e dal grafico 3.3 risulta plausibile l'ipotesi di normalità per la distribuzione dei residui. La tabella 3.4 evidenzia anche in questo caso un valore di mediana maggiore di zero, ma decisamente più basso rispetto al precedente. Abbiamo individuato la presenza di un *outlier*[2], infatti la percentuale che si riferisce ad ottobre 2009 è stata sovrastimata di molto: il dato reale è pari a 28.88%, mentre la stima risulta essere 29.56%. Osservando il grafico delle autocorrelazioni possiamo accettare l'ipotesi di casualità dei residui.

[2]L'*outlier* è stato individuato tramite il boxplot, che non riportiamo per non appesantire la presentazione dei risultati.

Tabella 3.4: *Summary* dei residui standardizzati - classe d'età 35-49 anni

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.712000	-0.548600	0.082110	0.0091096	0.723400	1.768000

Figura 3.3: Analisi dei residui: normalità e autocorrelazioni - classe d'età 35-49 anni

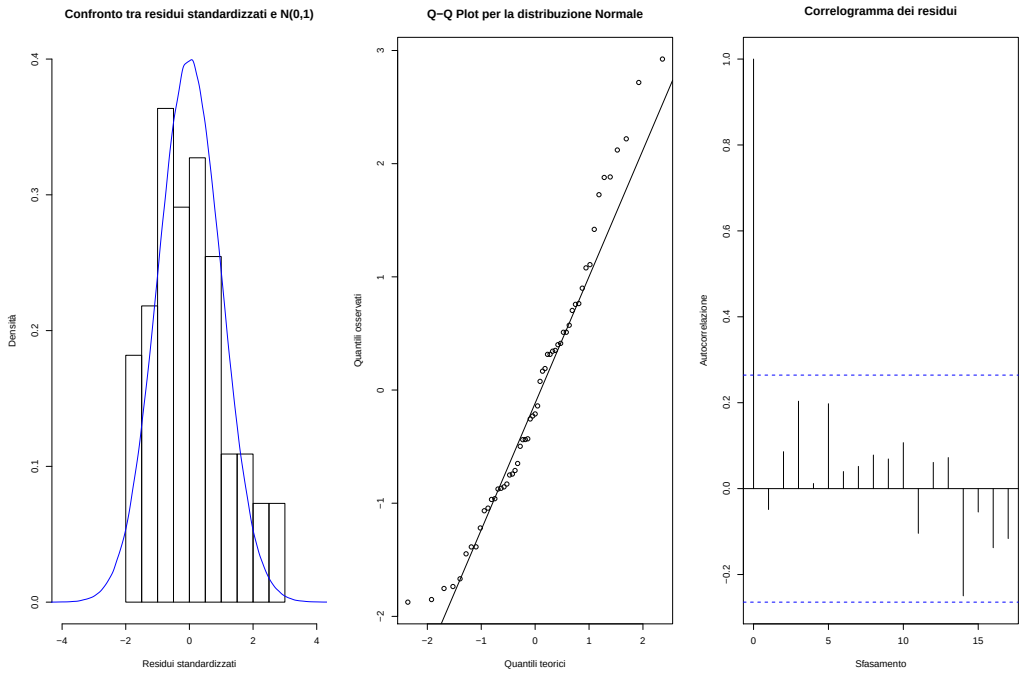


3. Classe d'età 50-69 anni: il *p-value* calcolato per il test di Shapiro-Wilk è superiore al livello di significatività considerato (0.05), ma non di molto, visto che è pari allo 0.07698 ed anche il grafico 3.4 e la tabella 3.5 denotano un'asimmetria verso sinistra e la presenza di alcuni valori che superano in maniera abbastanza marcata lo zero: nel complesso, dunque, la maggior parte delle percentuali risultano essere sovrastimate, mentre alcune presentano un problema di sottostima abbastanza rilevante. Come nei due casi precedenti, possiamo accettare l'ipotesi di aleatorietà dei residui, poichè il grafico delle autocorrelazioni non porta a pensare di aver tralasciato di modellare comportamenti particolari.

Tabella 3.5: *Summary* dei residui standardizzati - classe d'età 35-49 anni

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.87400	-0.87140	-0.21120	-0.01668	0.63790	2.92400

Figura 3.4: Analisi dei residui: normalità e autocorrelazioni - classe d'età 35-49 anni



Se proviamo a costruire modelli aventi componente di trend polinomiale di secondo o terzo grado, ci accorgiamo che i parametri relativi ai regressori temporali risultano non significativamente diversi da zero e che il BIC risulta più elevato di quello calcolato per il modello appena presentato (tabella 3.42). Di conseguenza, possiamo omettere i risultati ottenuti con la regressione quadratica e di terzo grado e scegliere come modello rappresentativo quello costruito ipotizzando la linearità della componente di trend.

3.3.1.2 Variabile di stratificazione: sesso

Suddividiamo il campione di riferimento in base al sesso: di media, mensilmente il campione rilevato è composto per il 49.25% da uomini, per il 50.75% da donne.

Tabella 3.6: Confronto tra BIC - partizionamento per classe d'età

	Grado 1	Grado 2	Grado 3
BIC	177.6	195.4	220.7

Costruiamo un modello con componente di trend lineare; le stime dei parametri "fissi" sono contenute nella tabella 3.7. La stima del coefficiente angolare medio risulta essere negativa e la sua varianza ridotta, rispetto al modello presentato in precedenza per il partizionamento secondo la classe d'età: ci aspettiamo una differenza minore tra le "velocità" di decrescita dei due gruppi. La correlazione tra μ_α e μ_β è pari a 1: un aumento (diminuzione) del valore dell'intercetta è accompagnato da un proporzionale aumento (diminuzione) del valore del coefficiente angolare; ciò significa che il gruppo caratterizzato dal livello iniziale di percentuale di fumatori più alto è caratterizzato da una decrescita meno ripida rispetto all'altro, seppur di poco, visto il valore relativamente piccolo assunto da $\hat{\sigma^2_\beta}$.

Tabella 3.7: Stime dei parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per sesso

	Stima	Std. Error	t value	*p-value*	Significatività
μ_α	-0.8945655	0.1579022	-5.665	1.47e-08	***
μ_β	-0.0016565	0.0003284	-5.044	4.56e-07	***

σ^2_α	σ^2_β	$\rho\sigma_\alpha\sigma_\beta$	ρ
4.964513e-02	6.723930e-10	5.777632e-06	1

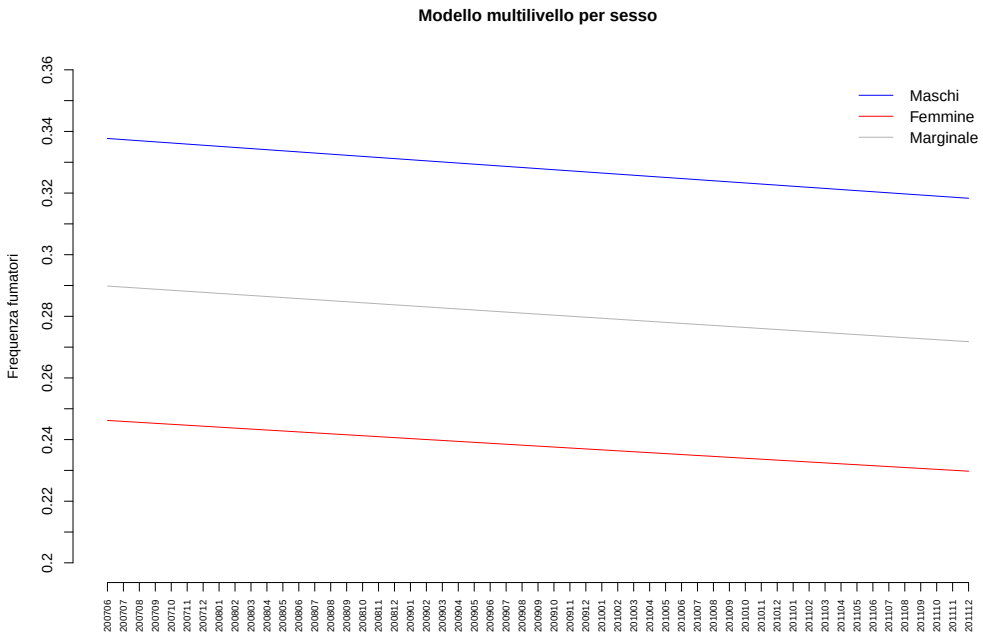
Gli scostamenti dal valor medio di ogni parametro relativi alle due sottopopolazioni, sono presentati nella tabella 3.8, mentre il grafico 3.5 illustra le rette di regressione stimate per il gruppo dei maschi e per il gruppo delle femmine, confrontate con la retta di regressione media. Le rette relative ai due gruppi sono quasi speculari rispetto alla retta costruita con le medie dei parametri; ciò era prevedibile, visto che siamo in presenza di una suddivisione del campione di ri-

ferimento in due parti con numerosità quasi uguali. La percentuale di fumatori tra gli uomini si sta riducendo, passando da una stima pari al 33.77% di giugno 2007, al 31.83% relativo a dicembre 2011. Le donne registrano un calo dal 24.62% al 22.97%.

Tabella 3.8: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per sesso

Gruppo	η_j	ζ_j
Maschi	0.2227027	2.591783e-05
Femmine	-0.2226765	-2.591479e-05

Figura 3.5: Rappresentazione del modello multilivello (con componente di trend lineare) - partizionamento per sesso



Concentriamoci sull'analisi dei residui, effettuata dopo un'opportuna standardizzazione di essi.

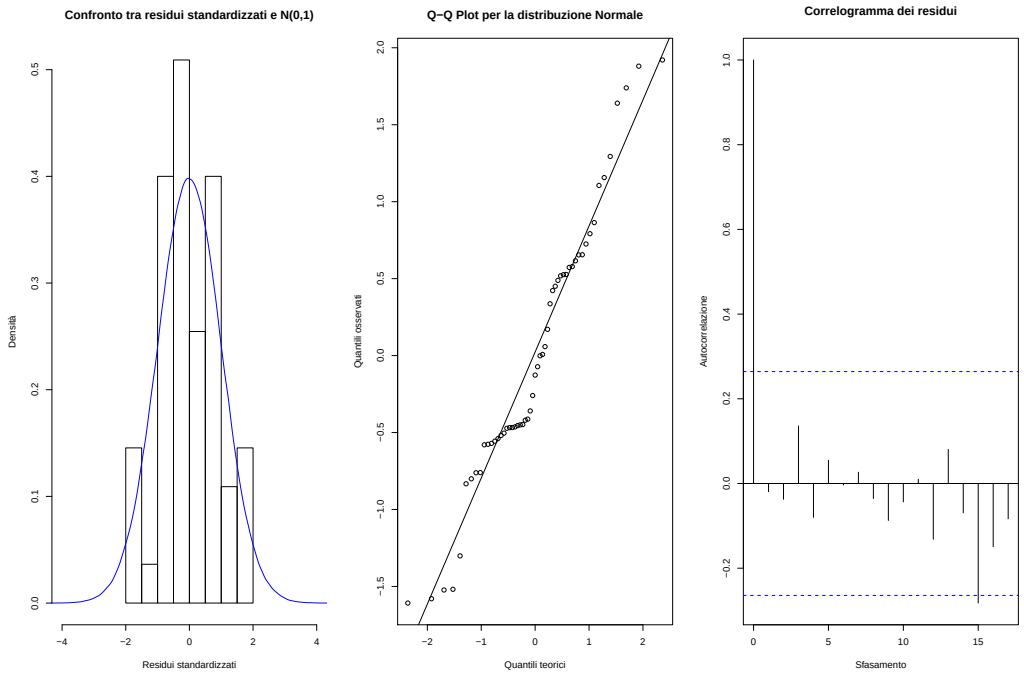
1. Sesso maschile: il *p-value* del test di Shapiro-Wilk ci porta a non rifiutare l'ipotesi di normalità, ma non si discosta così tanto dal limite di 0.05, essendo

pari a 0.07037. Osservando il grafico 3.6 e la tabella 3.3 sembra esserci una modesta asimmetria verso sinistra: il 55% dei valori risultano essere delle sovrastime, inoltre vi sono andamenti irregolari, nonostante la mancanza di valori anomali. Risulta essere significativo il coefficiente di autocorrelazione a livello $k = 15$, ma possiamo confermare l'ipotesi di casualità degli errori perchè, scelto un livello di significatività pari allo 0.05, non vi sarebbe nulla di strano se, anche per dati generati da un processo stocastico *white noise*, fra i primi 20 coefficienti di autocorrelazione campionari, uno risultasse significativo [Di Fonzo, Lisi 2005] .

Tabella 3.9: *Summary* dei residui standardizzati - sesso maschile

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.608000	-0.529000	-0.126600	0.005379	0.575700	1.921000

Figura 3.6: Analisi dei residui: normalità e autocorrelazioni - sesso maschile



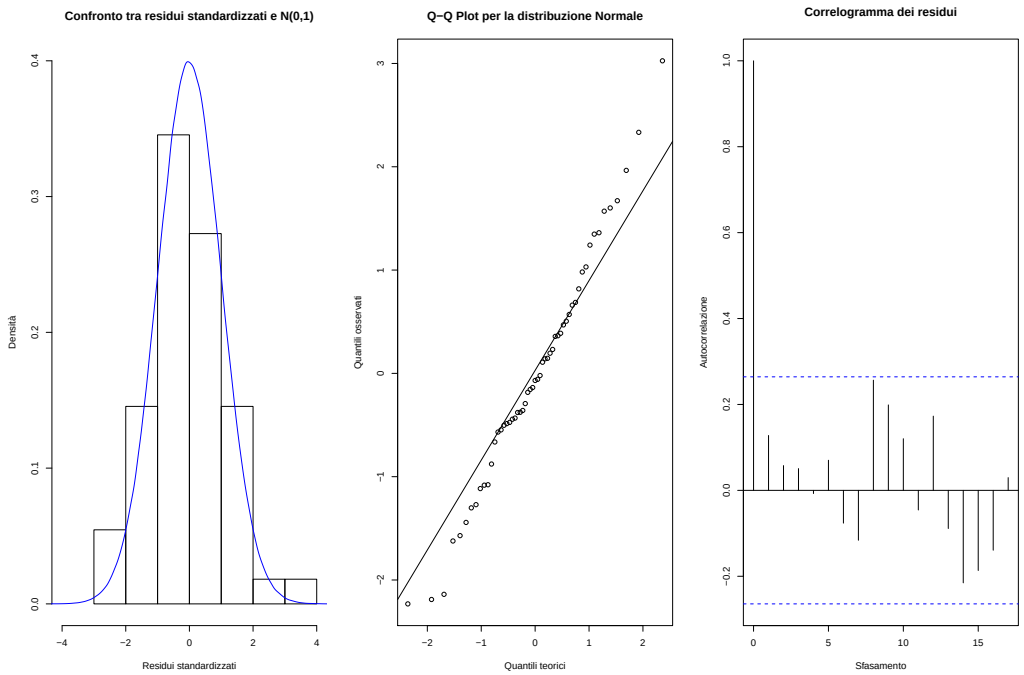
2. Sesso femminile: l'ipotesi di normalità dei residui è da accettare, visto il valore piuttosto alto assunto dal *p-value* del test di Shapiro-Wilk (0.786). Il

grafico 3.7 non presenta particolari irregolarità e la mediana (tabella 3.10) non si discosta molto dallo zero, ad indicare un'asimmetria trascurabile; da notare solo la presenza di un *outlier* che segnala una sottostima di non poco conto, in corrispondenza di febbraio 2008: la percentuale stimata di donne fumatrici è del 24.37%, mentre il dato reale è pari a 27.01%. L'ipotesi di aleatorietà è confermata dall'assenza di ρ_k significativamente diversi da zero.

Tabella 3.10: *Summary* dei residui standardizzati - sesso femminile

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.231000	-0.557400	-0.068590	-0.005379	0.615500	3.026000

Figura 3.7: Analisi dei residui: normalità e autocorrelazioni - sesso femminile



Anche in questa circostanza, sembra preferibile una componente di trend lineare: i valori del BIC (tabella 3.11) ci informano del fatto che l'aumento del numero di regressori provoca un aumento della variabilità spiegata non sufficiente a compensare la perdita di un grado di libertà.

Tabella 3.11: Confronto tra BIC - partizionamento per sesso

	Grado 1	Grado 2	Grado 3
BIC	117.0	129.9	175.5

3.3.1.3 Variabile di stratificazione: livello d'istruzione

Se consideriamo come criterio di suddivisione il livello di istruzione, vengono individuate quattro sottopopolazioni: coloro che non possiedono alcun titolo di studio oppure hanno conseguito la licenza elementare (che costituiscono mensilmente, di media, il 12% circa del campione), coloro che hanno conseguito la licenza di scuola media inferiore (circa il 31%), coloro che possiedono il diploma di scuola media inferiore (quasi il 44%) ed infine quelli che hanno un diploma universitario o una laurea (quasi il 13%).

Le stime delle medie e degli elementi della matrice varianze-covarianze che caratterizzano le distribuzioni dei parametri sono riportate nella tabella 3.12: il coefficiente angolare medio indica un andamento decrescente del fenomeno, com'era prevedibile supporre, visti i risultati delle precedenti analisi; presenta una varianza non particolarmente elevata, dell'ordine di quella stimata per i β_j relativi al partizionamento per sesso. La correlazione tra i parametri è pari a -1, quindi, come nel caso di suddivisione per classe d'età, un aumento dell'intercetta è accompagnato da una proporzionale diminuzione del coefficiente angolare, determinando così una decrescita più rapida. Il fenomeno rilevato all'interno di un gruppo di individui caratterizzato da una percentuale iniziale più bassa di fumatori, presenta invece una tendenza di fondo più costante (o eventualmente, crescente).

Gli scostamenti dai valori medi sono contenuti nella tabella 3.13 e sono visibili nel grafico 3.8: I due gruppi meno numerosi, ossia quelli comprendenti coloro che non possiedono alcun titolo di studio o hanno la licenza elementare, e coloro

Tabella 3.12: Stime dei parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per livello di'istruzione

	Stima	Std. Error	t value	*p-value*	Significatività
μ_α	-0.9958646	0.1256371	-7.927	2.25e-15	***
μ_β	-0.0018768	0.0003286	-5.712	1.12e-08	*

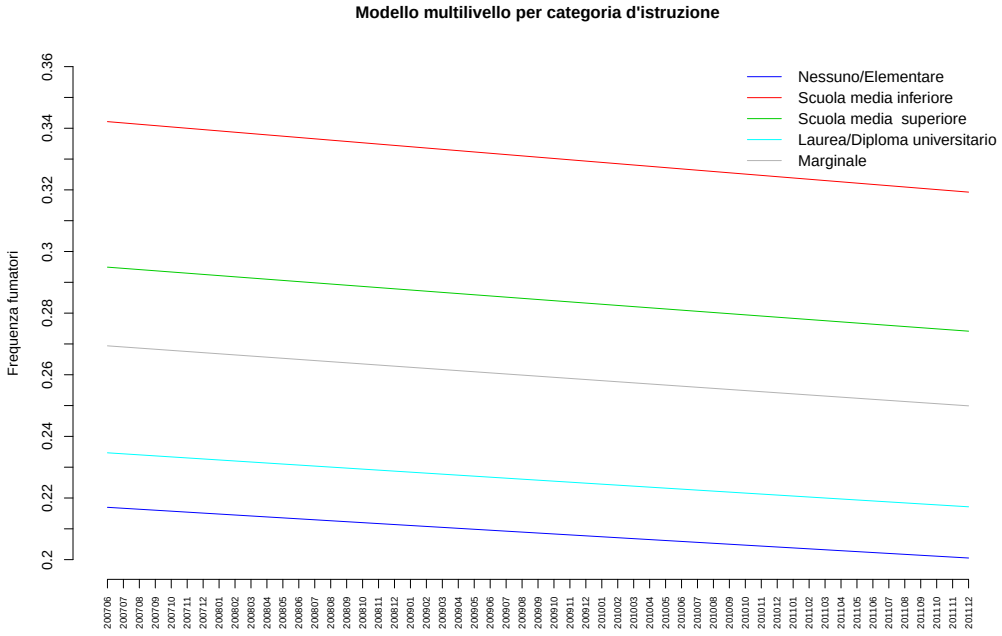
σ^2_α	σ^2_β	$\rho\sigma_\alpha\sigma_\beta$	ρ
6.264793e-02	7.672174e-10	-6.932863e-06	-1

che hanno conseguito il diploma universitario o la laurea, sono caratterizzati da una percentuale di fumatori più bassa della media, soprattutto la prima sottopopolazione. I due gruppi intermedi presentano invece una percentuale di fumatori più alta rispetto alla media, in particolar modo tra coloro che possiedono la licenza di scuola media inferiore, tra cui comunque la decrescita del fenomeno è un po' più marcata.

Tabella 3.13: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione

Gruppo	η_j	ζ_j
Nessuno/elementare	-0.2856463	3.161072e-05
Media inferiore	0.3440603	-3.807505e-05
Media superiore	0.1261602	-1.396137e-05
Laurea	-0.1843978	2.040618e-05

Figura 3.8: Rappresentazione del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione



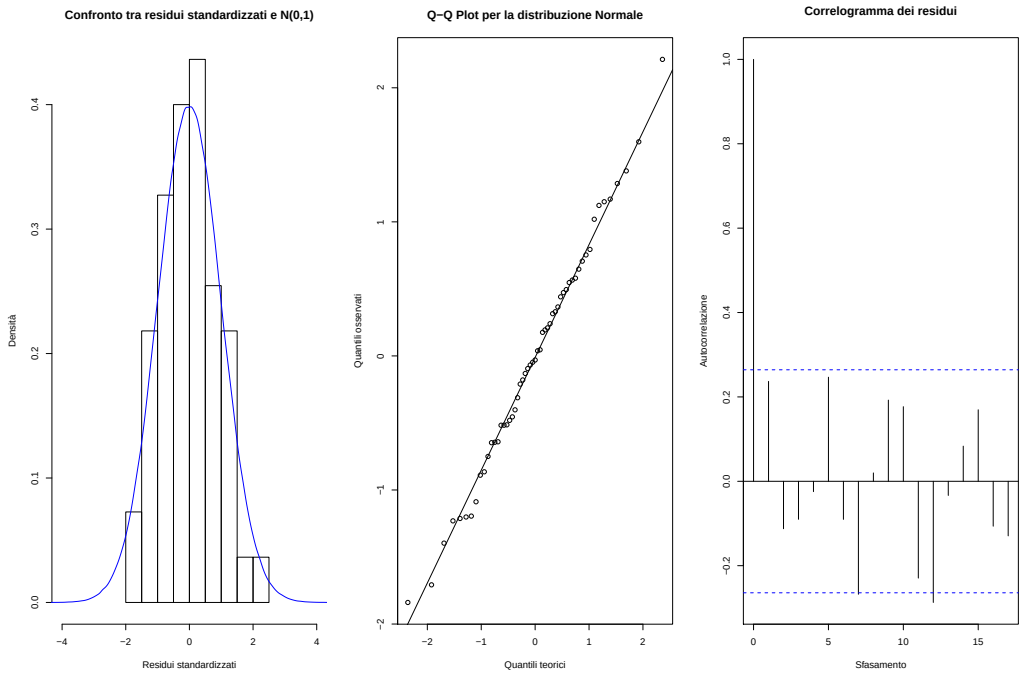
Passiamo ad analizzare i residui standardizzati.

1. Nessun titolo di studio/licenza elementare: il *p-value* del test di Shapiro-Wilk (0.9897), i dati contenuti nella tabella 3.14 ed il grafico 3.9 indicano che l'ipotesi di normalità dei residui è da accettare, poichè non si notano asimmetria o irregolarità evidenti e non vi sono nemmeno valori anomali. L'analisi del correlogramma ci porta a considerare valida l'ipotesi di casualità.

Tabella 3.14: *Summary* dei residui standardizzati - nessun titolo di studio/licenza elementare

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.839000	-0.578800	-0.030330	-0.007472	0.556400	2.212000

Figura 3.9: Analisi dei residui: normalità e autocorrelazioni - nessun titolo di studio/licenza elementare

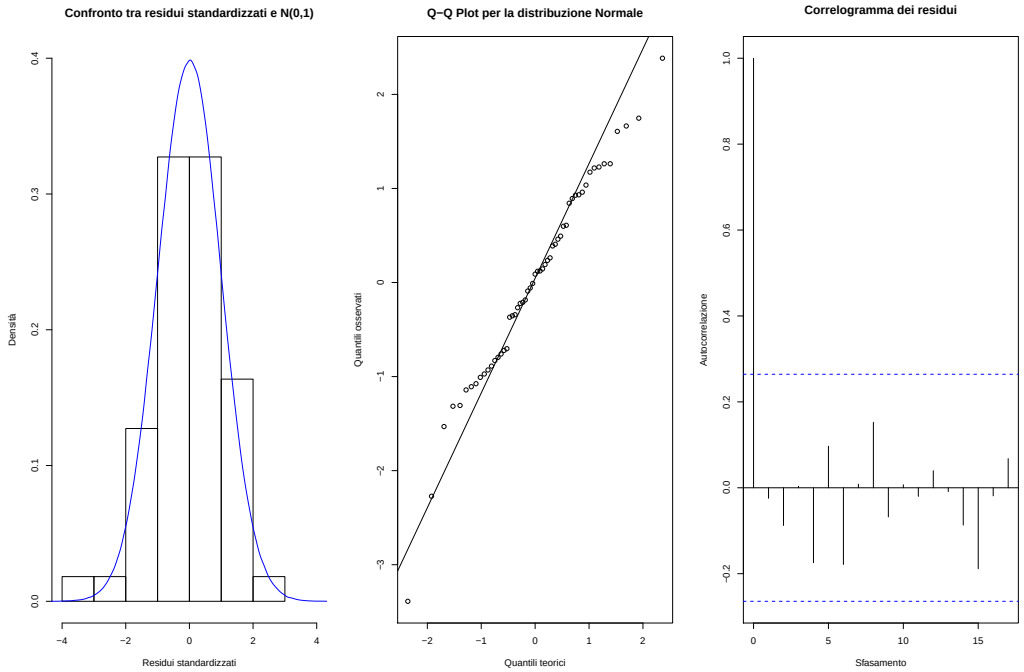


2. Licenza scuola media inferiore: l'ipotesi di normalità non va rifiutata, poichè il test di Shapiro-Wilk presenta un *p-value* pari a 0.4719. Non si notano particolari asimmetrie (tabella 3.15 e grafico 3.10). C'è invece un *outlier*, un valore sovrastimato, in corrispondenza di maggio 2011: la percentuale reale è pari al 27.84%, la sua stima invece è di 32.22%. L'ipotesi di aleatorietà è confermata, non sembrano esserci comportamenti non casuali non colti dal modello.

Tabella 3.15: *Summary* dei residui standardizzati - licenza scuola media inferiore

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.389000	-0.777600	0.089250	0.007362	0.868000	2.384000

Figura 3.10: Analisi dei residui: normalità e autocorrelazioni - licenza scuola media inferiore

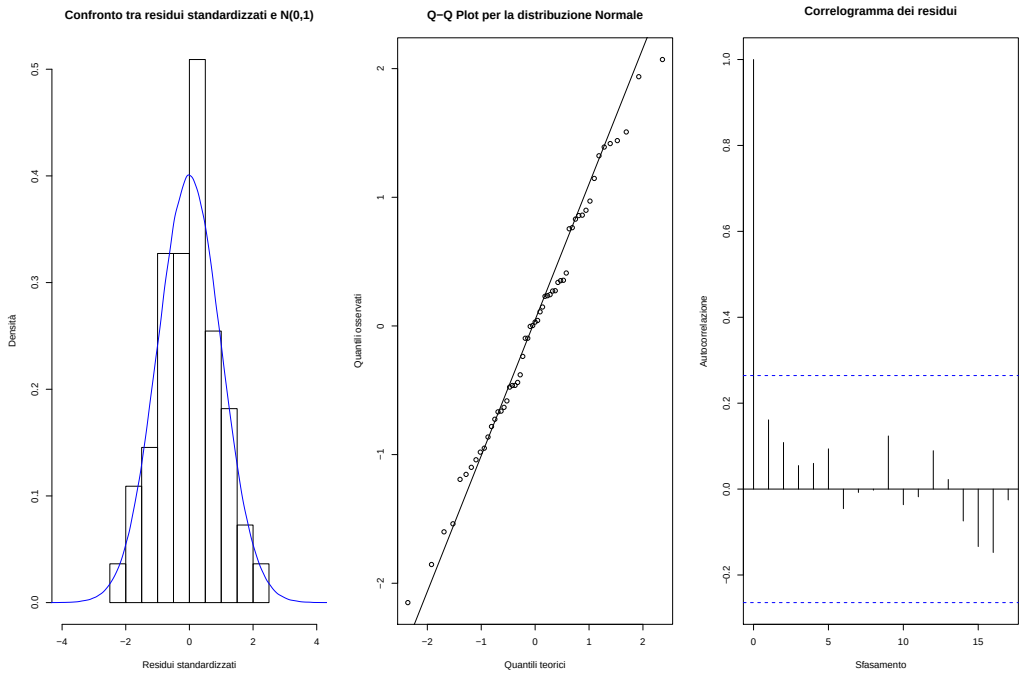


3. Diploma scuola media superiore: anche in questo caso il valore del *p-value* del test di Shapiro-Wilk è piuttosto elevato (0.9191) e nè dal grafico 3.11 nè dalla tabella 3.16 riscontriamo irregolarità evidenti; l'ipotesi di normalità è dunque accettata ed anche quella di casualità, come confermato dal grafico delle autocorrelazioni.

Tabella 3.16: *Summary* dei residui standardizzati - diploma scuola media superiore

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.15000	-0.66490	0.03013	0.00163	0.75960	2.07100

Figura 3.11: Analisi dei residui: normalità e autocorrelazioni - diploma scuola media superiore

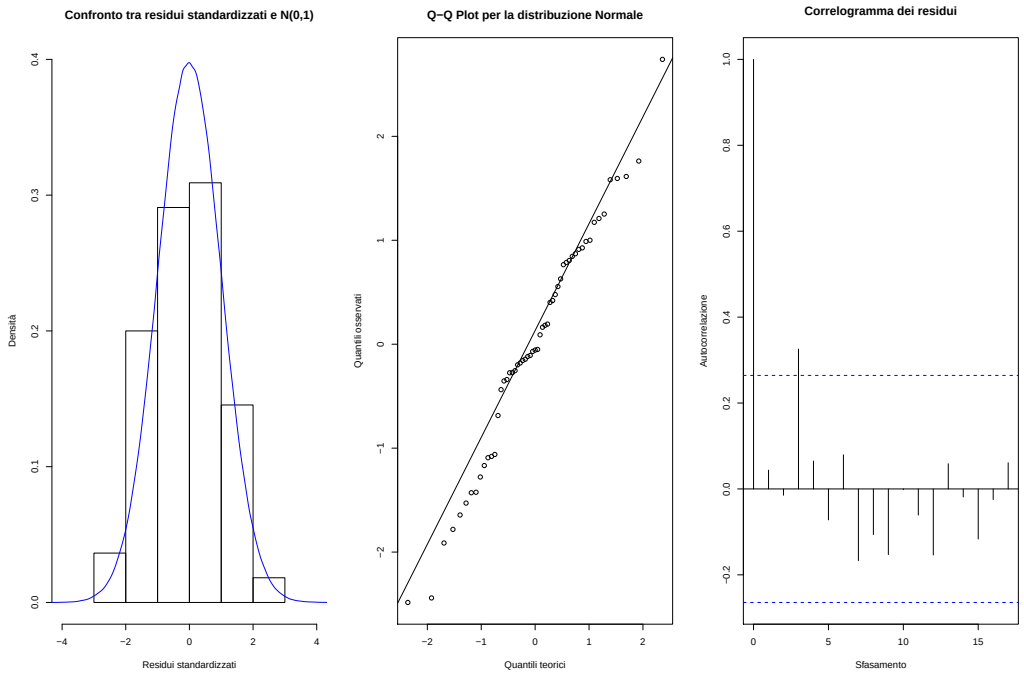


4. Diploma universitario/laurea: *p-value* pari a 0.5778 per il test di Shapiro-Wilk, dunque l'ipotesi di normalità non può essere rifiutata. Dal valore della mediana (tabella 3.17) e dal grafico 3.12 si nota però una leggera asimmetria verso sinistra: circa il 53% delle percentuali risulta essere sovrastimata, seppur non in modo anomalo, mentre la perfetta simmetria si avrebbe con il 50%. Per quanto riguarda le autocorrelazioni, risulta significativo il coefficiente in corrispondenza di $k = 3$, ma, come spiegato a pagina 95, non rifiutiamo l'ipotesi di aleatorietà.

Tabella 3.17: *Summary* dei residui standardizzati - diploma universitario/laurea

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.48500	-0.56190	-0.05648	-0.00152	0.82530	2.74000

Figura 3.12: Analisi dei residui: normalità e autocorrelazioni - diploma universitario/laurea



Anche in questo caso, il modello migliore risulta essere quello con componente di trend lineare, come confermato dai valori assunti dal BIC (tabella 3.18).

Tabella 3.18: Confronto tra BIC - partizionamento per livello d'istruzione

	Grado 1	Grado 2	Grado 3
BIC	235.7	253.9	309.6

3.3.1.4 Variabile di stratificazione: regione

Considerando la regione d'appartenenza come criterio per partizionare il campione di riferimento, otteniamo 21 sottopopolazioni. Non dimentichiamo che seguendo questo approccio ci imbattiamo in alcuni dati mancanti, come spiegato a pagina 54.

Come nei casi precedenti, presentiamo le stime del modello costruito con componente di trend lineare e facciamo alcune valutazioni sui valori contenuti nella

tabella 3.19: di media il trend è decrescente e la varianza relativa al coefficiente angolare è più alta di quelle riscontrate nei casi precedenti, probabilmente perchè in presenza di un numero più alto di gruppi è facile aspettarsi maggiore variabilità. Tra i due parametri considerati c'è una ridotta correlazione negativa.

Tabella 3.19: Stime dei parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per regione

	Stima	Std. Error	t value	*p-value*	Significatività
μ_α	-0.8961078	0.0259435	-34.54	< 2e-16	***
μ_β	-0.0016381	0.0005214	-3.14	0.00168	**

σ^2_α	σ^2_β	$\rho\sigma_\alpha\sigma_\beta$	ρ
1.051254e-02	2.621230e-06	-2.821534e-05	-0.1699726

La tabella 3.20 contiene gli scostamenti dalle medie dei due parametri, analizzando i quali possiamo valutare l'andamento del fenomeno del fumo all'interno di ogni regione. Di indubbia utilità è anche il grafico 3.13, che illustra le rette di regressione stimate.

- Particolare attenzione meritano le rette che modellano l'andamento del fumo in Umbria ed in Abruzzo: in queste due regioni la percentuale di fumatori è in aumento, mentre di media è stato riscontrato un andamento decrescente, rilevato anche in tutti i gruppi risultanti dai partizionamenti effettuati in precedenza.

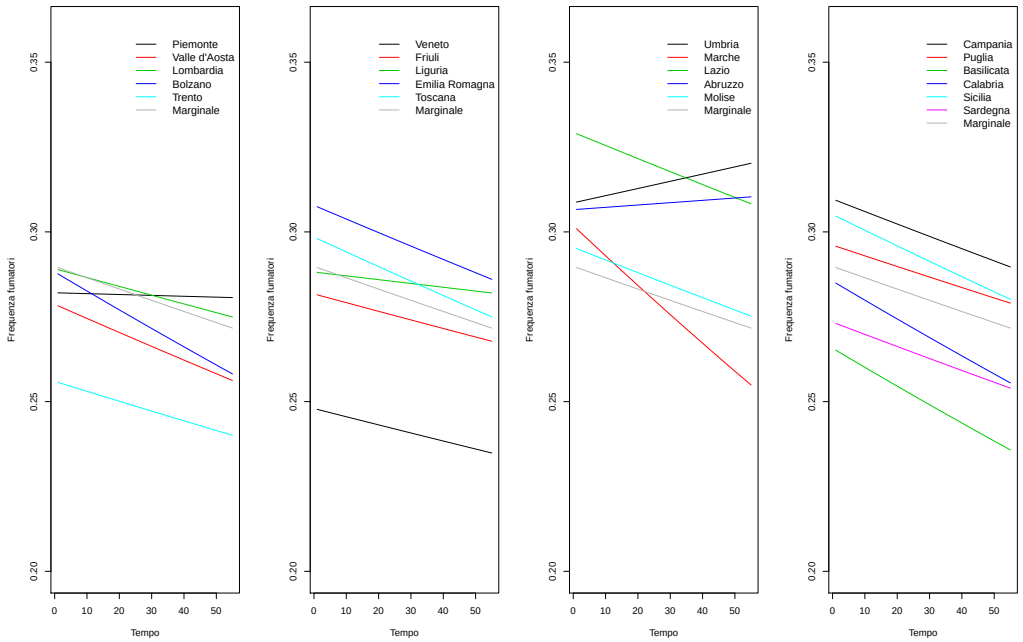
- La provincia autonoma di Trento, il Veneto e la Basilicata sembrano essere le regioni in cui nel complesso si fuma meno.

- Le Marche sono la regione che presenta il maggior tasso di decrescita, mentre tra i piemontesi la percentuale di fumatori nel corso del tempo sembra rimanere pressochè costante.

Tabella 3.20: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per regione

Gruppo	η_j	ζ_j
Piemonte	-0.038092236	1.509807e-03
Valle d'Aosta	-0.054967107	-4.456939e-04
Lombardia	-0.003313308	3.597378e-04
Bolzano	-0.008028059	-1.115921e-03
Trento	-0.170879574	8.685349e-05
Veneto	-0.213394700	3.373676e-04
Friuli Venezia Giulia	-0.039821137	3.663173e-04
Liguria	-0.008120205	1.087991e-03
Emilia Romagna	0.085778098	-2.601022e-04
Toscana	0.041473852	-4.513993e-04
Umbria	0.089321512	2.620631e-03
Marche	0.057620102	-2.619773e-03
Lazio	0.184927678	-1.256164e-04
Abruzzo	0.079836539	1.959632e-03
Molise	0.027198237	-1.719941e-04
Campania	0.094537573	-9.628254e-05
Puglia	0.030160925	1.209036e-04
Basilicata	-0.120285628	-1.265824e-03
Calabria	-0.021269707	-1.126525e-03
Sicilia	0.073097501	-5.590702e-04
Sardegna	-0.081301679	-1.817902e-04

Figura 3.13: Rappresentazione del modello multilivello (con componente di trend lineare) - partizionamento per regione



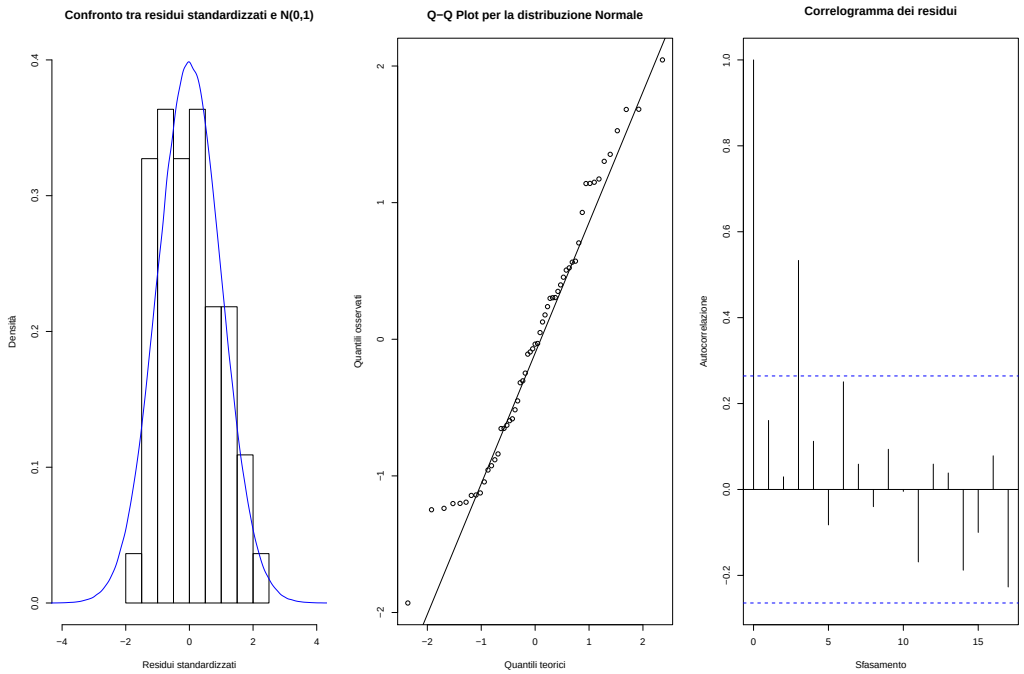
Effettuiamo l'analisi dei residui, dopo averli standardizzati.

1. Piemonte: il test di Shapiro-Wilk presenta un *p-value* pari a 0.2363, che ci fa accettare l'ipotesi di normalità, nonostante una lieve asimmetria verso sinistra (sovrastima) evidenziata dal grafico 3.14. Nella tabella 3.21 è riportato il *summary* dei residui. Dal grafico delle autocorrelazioni risulta essere particolarmente significativo ρ_3: forse abbiamo tralasciato qualche componente modellabile, o forse possiamo ricondurci alla spiegazione offerta a pagina 95.

Tabella 3.21: *Summary* dei residui standardizzati - Piemonte

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.93000	-0.74640	-0.03568	-0.01194	0.54350	2.04500

Figura 3.14: Analisi dei residui: normalità e autocorrelazioni - Piemonte



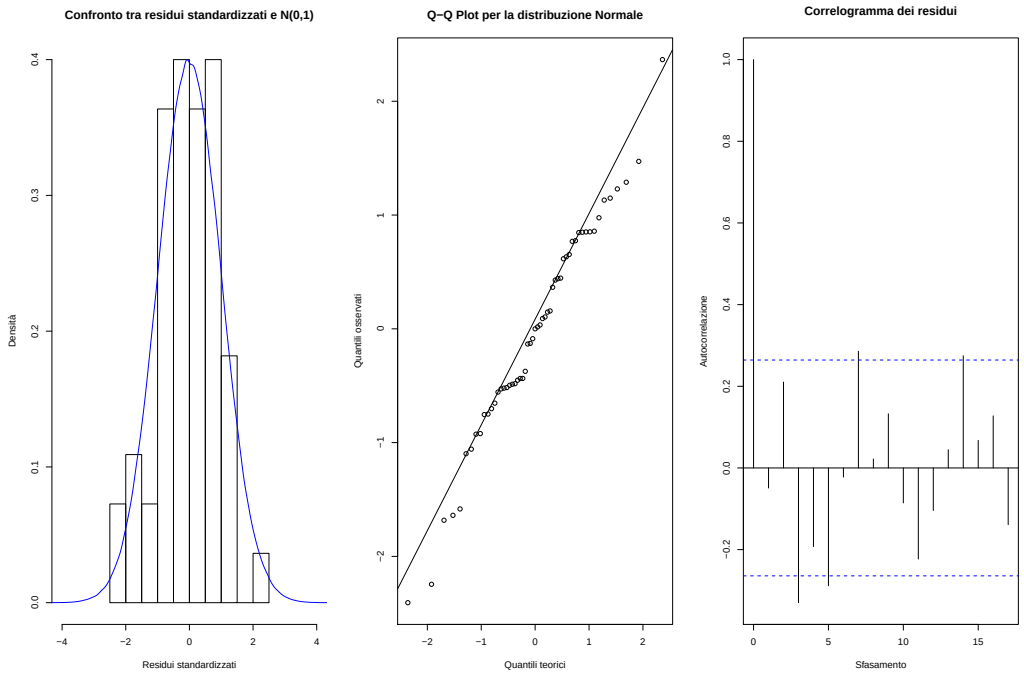
2. Valle d'Aosta: l'ipotesi di normalità dei dati è confermata dal *p-value* del test di Shapiro-Wilk (0.5901) e dal grafico 3.15, nonchè dai valori contenuti

nella tabella 3.22. Il grafico delle autocorrelazioni presenta invece alcuni coefficienti significativi: potremmo essere in dubbio sulla casualità vera e propria dei residui, nel senso che potremmo sospettare di aver tralasciato di modellare qualche componente; ma poichè l'ipotesi di normalità, che non viene rifiutata, include anche quella di aleatorietà, possiamo rifarci anche in questo caso a quanto detto a pagina 95.

Tabella 3.22: *Summary* dei residui standardizzati - Valle d'Aosta

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.4070000	-0.5422000	-0.0001096	-0.0455500	0.7111000	2.3660000

Figura 3.15: Analisi dei residui: normalità e autocorrelazioni - Valle d'Aosta



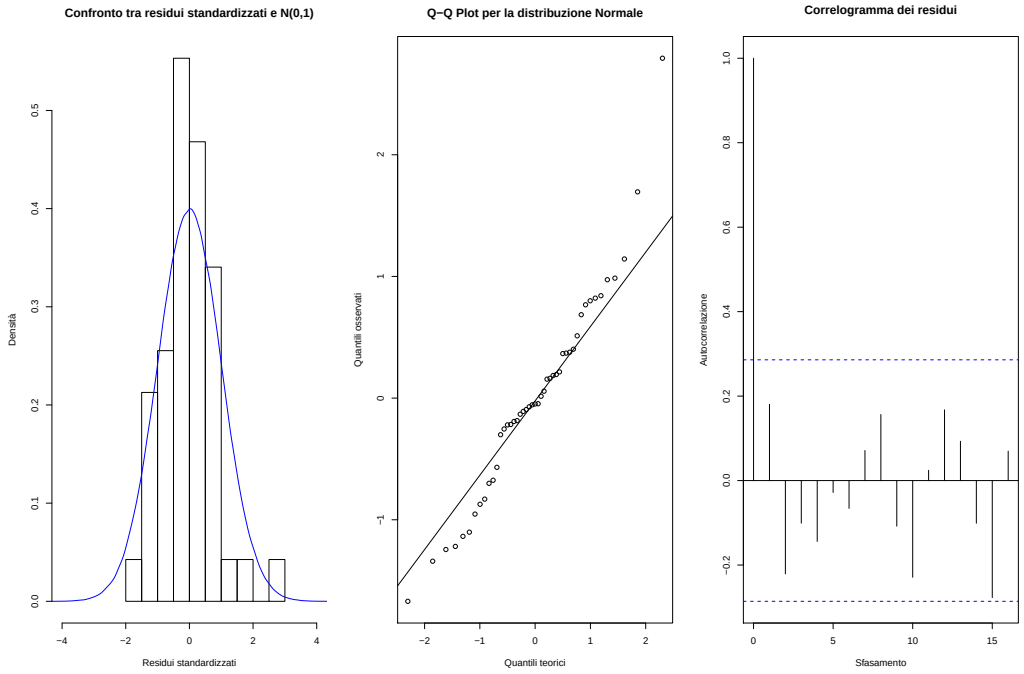
3. Lombardia: in questo caso il *p-value* risultante dal test di Shapiro-Wilk è pari allo 0.1223, un valore superiore al livello di significatività $\alpha = 0.05$ ma non di molto. Il grafico 3.16 e la tabella 3.23 sembrano confermare i nostri dubbi sull'effettiva normalità dei residui: vi è un'asimmetria verso sinistra, infatti i valori sovrastimati sono pari al 60%. Inoltre ci sono due

outlier, che si riferiscono a due sottostime piuttosto discrepanti dai dati reali, in corrispondenza di marzo 2010 e ottobre 2011: le percentuali rilevate sono rispettivamente pari a 34.68% e 38.28%, mentre le stime sono 28.03% e 27.54%. Il correlogramma non registra alcun coefficiente significativamente diverso da zero.

Tabella 3.23: *Summary* dei residui standardizzati - Lombardia

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.670000	-0.434900	-0.048280	0.005938	0.389400	2.793000

Figura 3.16: Analisi dei residui: normalità e autocorrelazioni - Lombardia



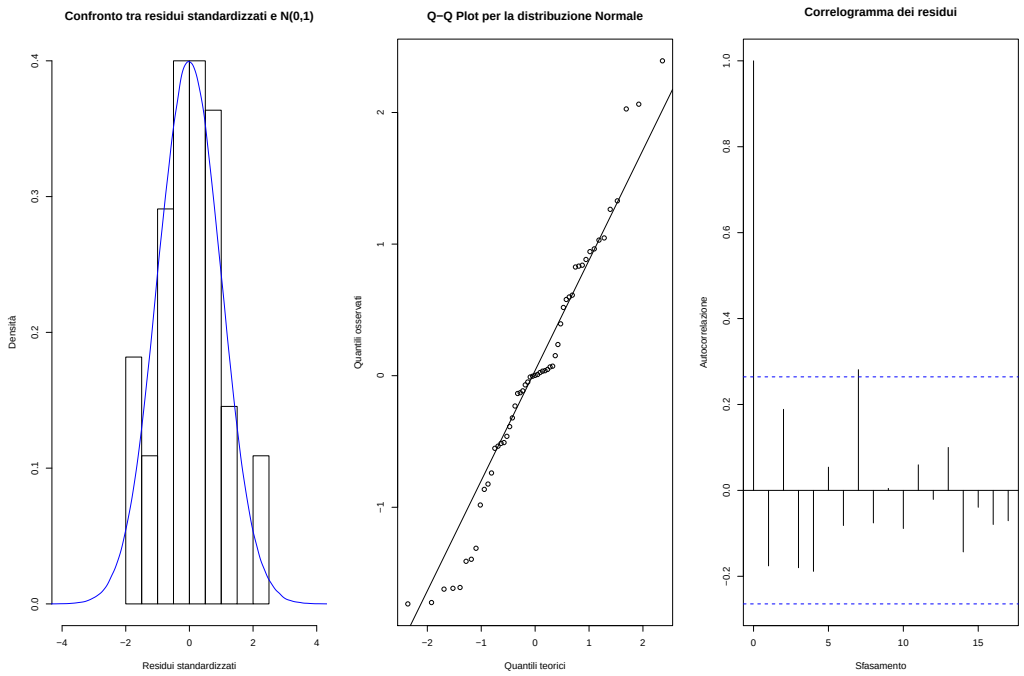
4. Bolzano: il test di Shapiro-Wilk presenta un *p-value* pari a 0.2049, portandoci all'accettazione dell'ipotesi di normalità. Dal grafico 3.17 non notiamo particolari asimmetrie, come confermato dai valori contenuti nella tabella 3.24, ma vi sono andamenti irregolari attorno al valor medio. Inoltre è presente un *outlier*: la stima della percentuale di fumatori relativa al mese di settembre 2009 è pari a 27.26% mentre il dato reale risulta essere 48%; la

discrepanza è notevole, ma ricordiamo (pagina 24) che la serie della frazione di fumatori di Bolzano presenta un'elevata variabilità, dovuta anche alla bassa numerosità dei campioni mensilmente osservati. Dal correlogramma rileviamo la significatività di un coefficiente, ma l'ipotesi di aleatorietà, come già spiegato in precedenza (pagina 95), non è da rifiutare.

Tabella 3.24: *Summary* dei residui standardizzati - Bolzano

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.7340000	-0.5250000	0.0025740	-0.0004919	0.6050000	2.3930000

Figura 3.17: Analisi dei residui: normalità e autocorrelazioni - Bolzano



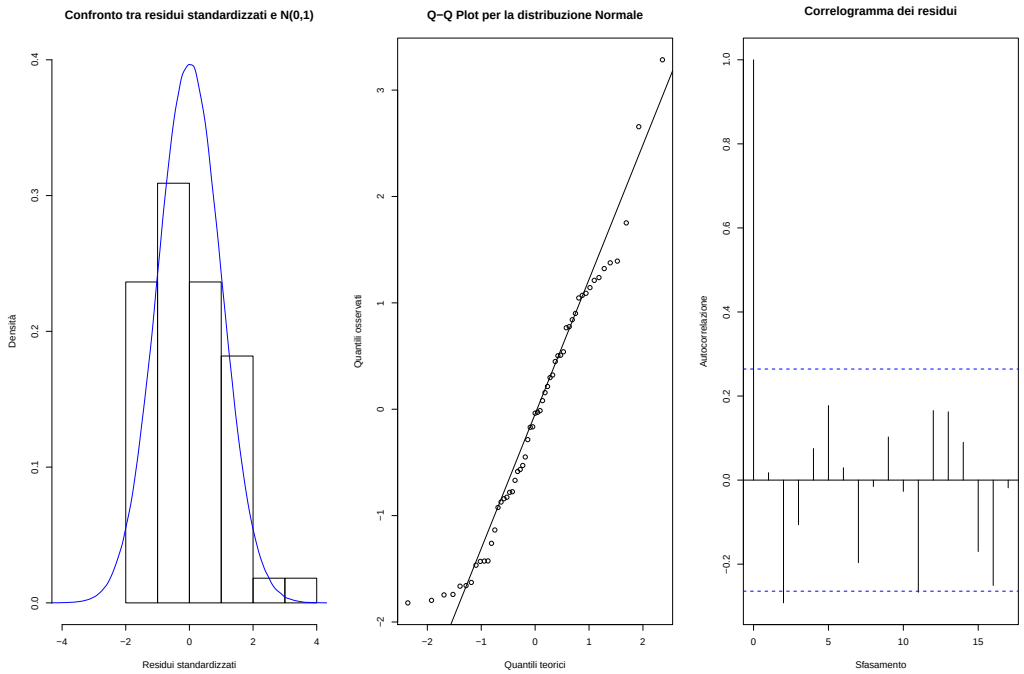
5. Trento: il valore del *p-value* ottenuto dal test di Shapiro-Wilk è molto vicino alla soglia di rifiuto dell'ipotesi nulla, ossia dell'ipotesi di normalità (0.06271); analizzando il grafico ed i valori dei residui (figura 3.18 e tabella 3.25) notiamo un'asimmetria verso sinistra, infatti circa il 55% delle frazioni calcolate dal modello risultano essere delle sovrastime. Dal correlogramma notiamo qualche coefficiente significativamente diverso da zero, o vicino

alla soglia della significatività, il che ci fa dubitare, oltre che dell'ipotesi di normalità, anche di quella di aleatorietà.

Tabella 3.25: *Summary* dei residui standardizzati - Trento

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.82000	-0.89790	-0.03721	-0.06873	0.80850	3.28500

Figura 3.18: Analisi dei residui: normalità e autocorrelazioni - Trento

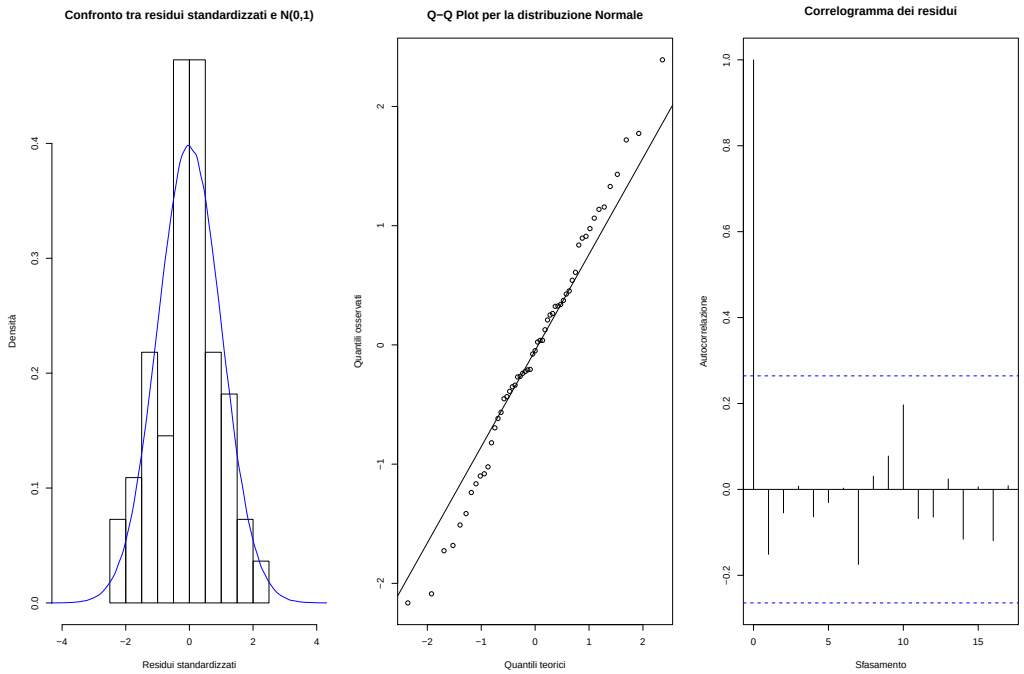


6. Veneto: in questo caso, il test di Shapiro-Wilk non lascia alcun dubbio sulla normalità degli errori, poichè presenta un *p-value* pari allo 0.9654; anche dal grafico 3.19 non notiamo irregolarità o asimmetrie. Rileviamo però la presenza di un valore anomalo (riassunto dei valori dei residui in tabella 3.26): il dato relativo ad aprile 2011 è sottostimato, poichè la percentuale rilevata è del 28.09%, mentre il modello ci offre un valore pari a 23.67%. Non risulta invece significativo alcun coefficiente di auocorrelazione.

Tabella 3.26: *Summary* dei residui standardizzati - Veneto

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.16400	-0.59130	-0.04880	-0.04418	0.49730	2.39100

Figura 3.19: Analisi dei residui: normalità e autocorrelazioni - Veneto

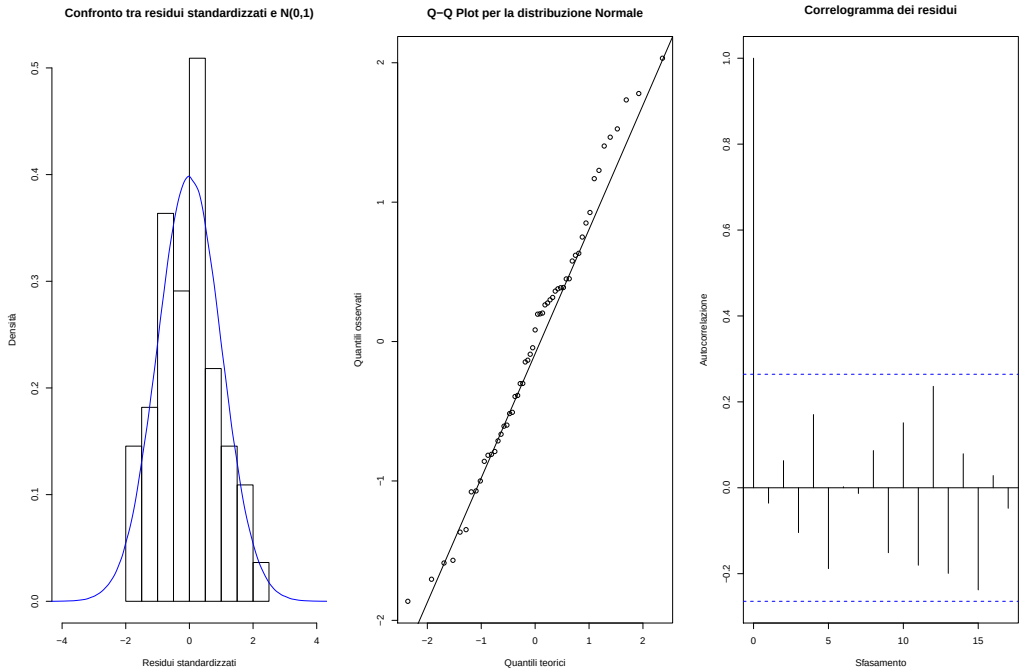


7. Friuli Venezia Giulia: anche per questa regione il risultato del test di Shapiro-Wilk è confortante, poichè il *p-value* è pari allo 0.6814. Inoltre, non si notano particolari asimmetrie (grafico 3.20 e tabella 3.27) e dal boxplot non risultano esserci valori anomali. Anche il grafico delle autocorrelazioni conferma l'ipotesi di casualità dei residui.

Tabella 3.27: *Summary* dei residui standardizzati - Friuli Venezia Giulia

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.863000	-0.688600	0.083600	-0.005935	0.514300	2.033000

Figura 3.20: Analisi dei residui: normalità e autocorrelazioni - Friuli Venezia Giulia

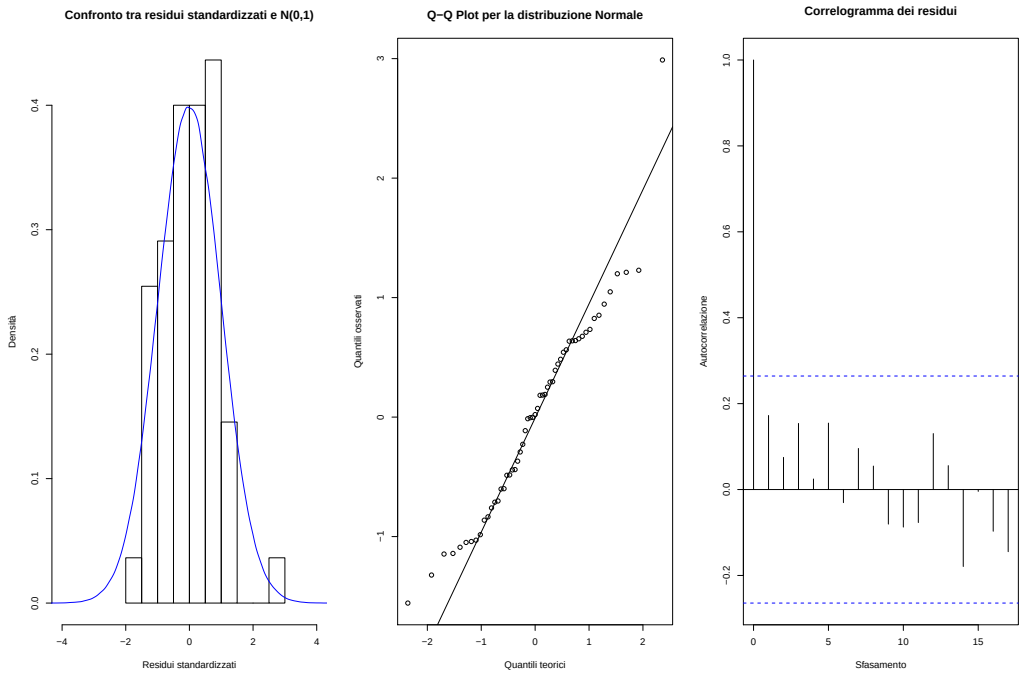


8. Liguria: completamente diversa è la situazione per quanto riguarda l'analisi dei residui tra valori reali e stimati della Liguria. L'ipotesi di casualità è confermata, mentre quella di normalità va rifiutata: il *p-value* del test di Shapiro-Wilk è pari allo 0.03771 e dal grafico si nota un andamento irregolare dei residui. Nel complesso non vi sono asimmetrie rilevanti, ma i residui non seguono una distribuzione Normale, come si può notare anche a vista d'occhio (figura 3.21 e tabella 3.28). Inoltre, il valore relativo a maggio 2011 risulta essere decisamente sottostimato dal modello: il dato osservato è pari al 40%, quello stimato è del 28.28%.

Tabella 3.28: *Summary* dei residui standardizzati - Liguria

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.55600	-0.65200	0.02091	0.01089	0.63690	2.98900

Figura 3.21: Analisi dei residui: normalità e autocorrelazioni - Liguria

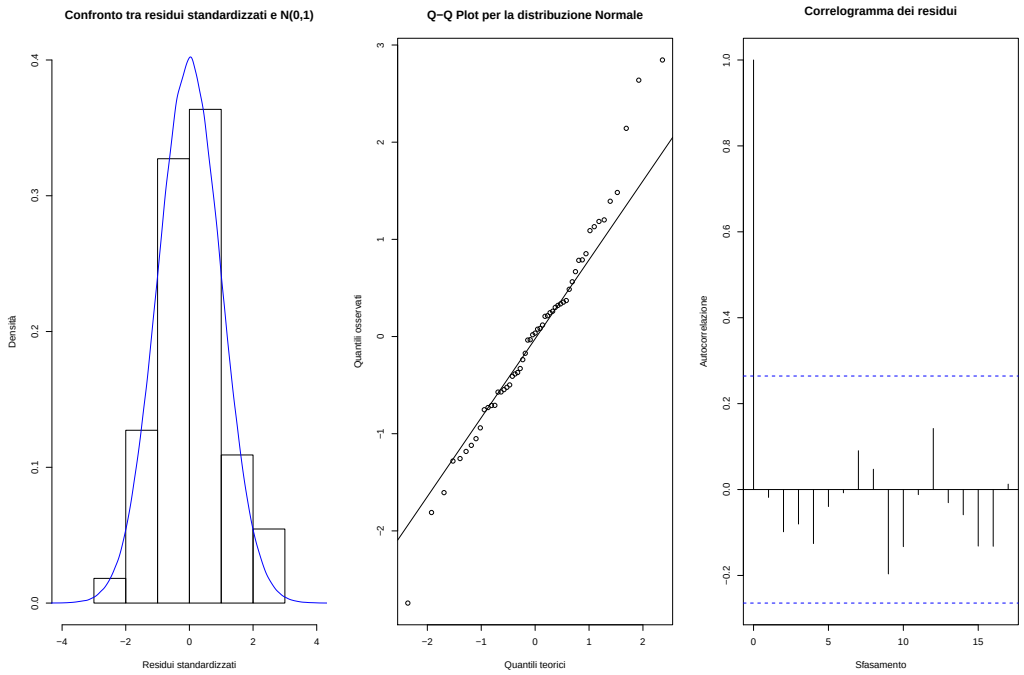


9. Emilia Romagna: l'ipotesi di normalità, secondo il test di Shapiro-Wilk, non vi rifiutata (*p-value*=0.4432). Dal grafico e dai valori assunti dai residui (figura 3.22 e tabella 3.29) risultano esserci però degli *outlier*, in corrispondenza di gennaio 2009, aprile 2009 e marzo 2011: nel secondo caso si tratta di una sovrastima, poichè il dato reale è pari a 23.54%, mentre quello stimato è 29.86%; gli altri due sono casi di sottostima perchè a gennaio 2009 è stata rilevata una frazione pari al 36.53% mentre la la stima è del 29.98%, il dato di marzo 2011 è 35.64% mentre il modello stima una percentuale pari al 28.95%. L'ipotesi di aleatorietà è confermata.

Tabella 3.29: *Summary* dei residui standardizzati - Emilia Romagna

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.74300	-0.57100	0.03495	0.02938	0.52480	2.84600

Figura 3.22: Analisi dei residui: normalità e autocorrelazioni - Emilia Romagna

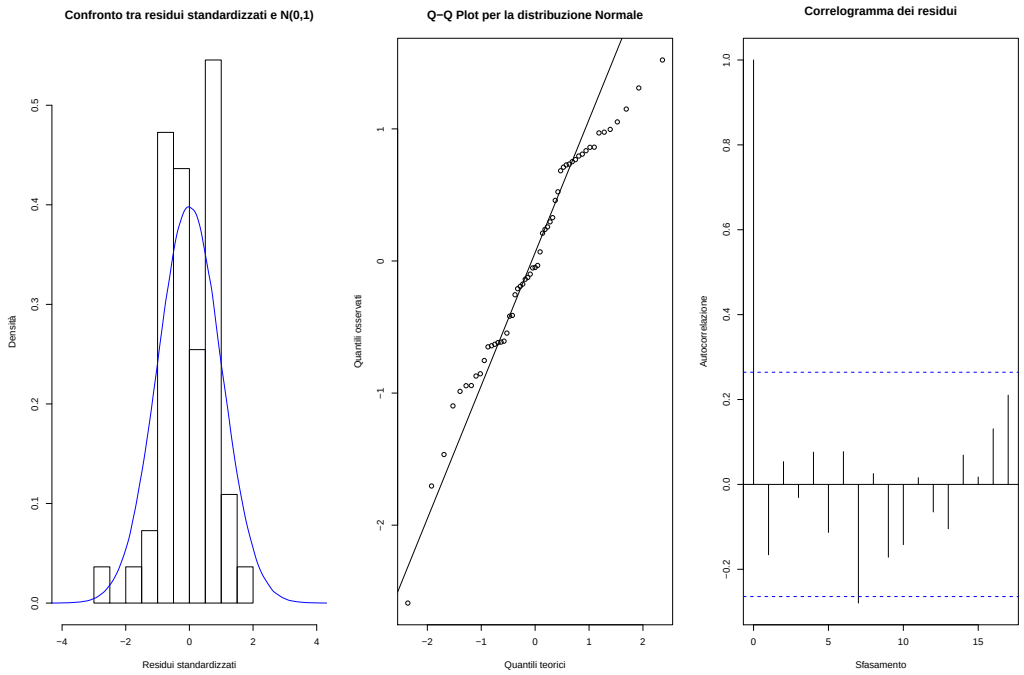


10. Toscana: il *p-value* ottenuto dal test di Shapiro-Wilk non è particolarmente alto, ma tale da permettere il non rifiuto dell'ipotesi di normalità (0.1003). Ciò è dovuto ad una discreta irregolarità nella distribuzione dei residui attorno al valor medio e ad una modesta asimmetria verso sinistra: quasi il 53% dei valori risulta essere sottostimato (grafico 3.23 e tabella 3.30). L'ipotesi di casualità è confermata dal grafico delle autocorrelazioni.

Tabella 3.30: *Summary* dei residui standardizzati - Toscana

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.590000	-0.616000	-0.049180	0.003898	0.742800	1.522000

Figura 3.23: Analisi dei residui: normalità e autocorrelazioni - Toscana

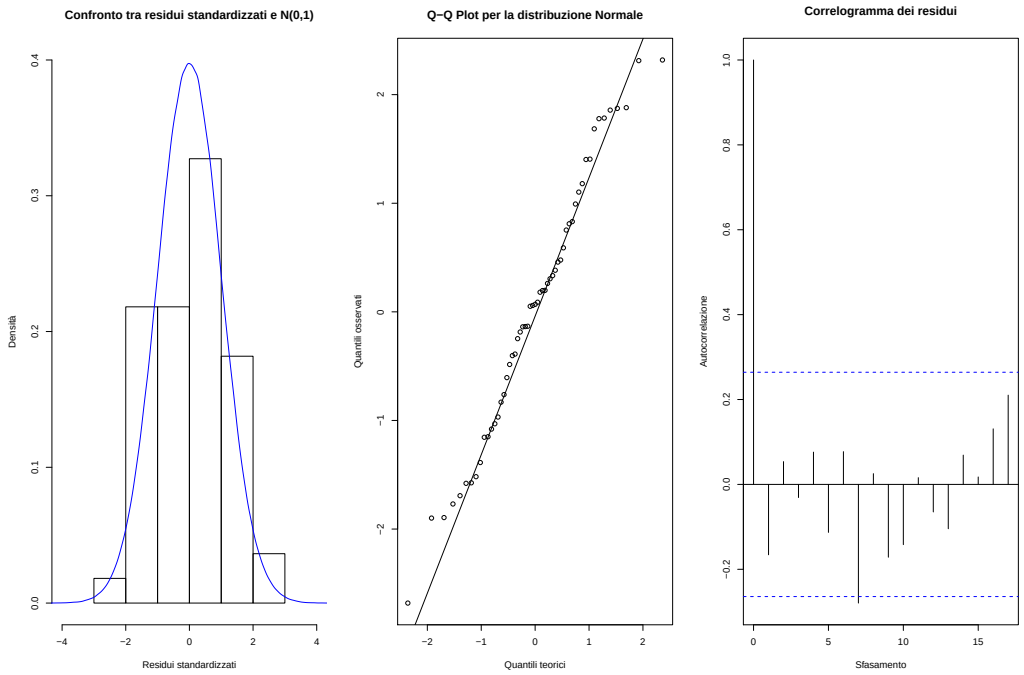


11. Umbria: il test di Shapiro-Wilk presenta un *p-value* pari a 0.482, quindi l'ipotesi di normalità è da accettare. Il grafico 3.24 non denota irregolarità attorno alla media ma, analizzando anche i valori dei residui, riassunti nella tabella 3.31, si rileva un'asimmetria verso destra: il 54.5% delle stime è minore dei dati osservati. Solo un coefficiente di autocorrelazione è significativamente diverso da zero, ma possiamo ricondurci alla spiegazione esposta a pag 95 e non rifiutare l'ipotesi di casualità.

Tabella 3.31: *Summary* dei residui standardizzati - Umbria

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.68100	-0.90000	0.06941	0.03513	0.82110	2.31900

Figura 3.24: Analisi dei residui: normalità e autocorrelazioni - Umbria

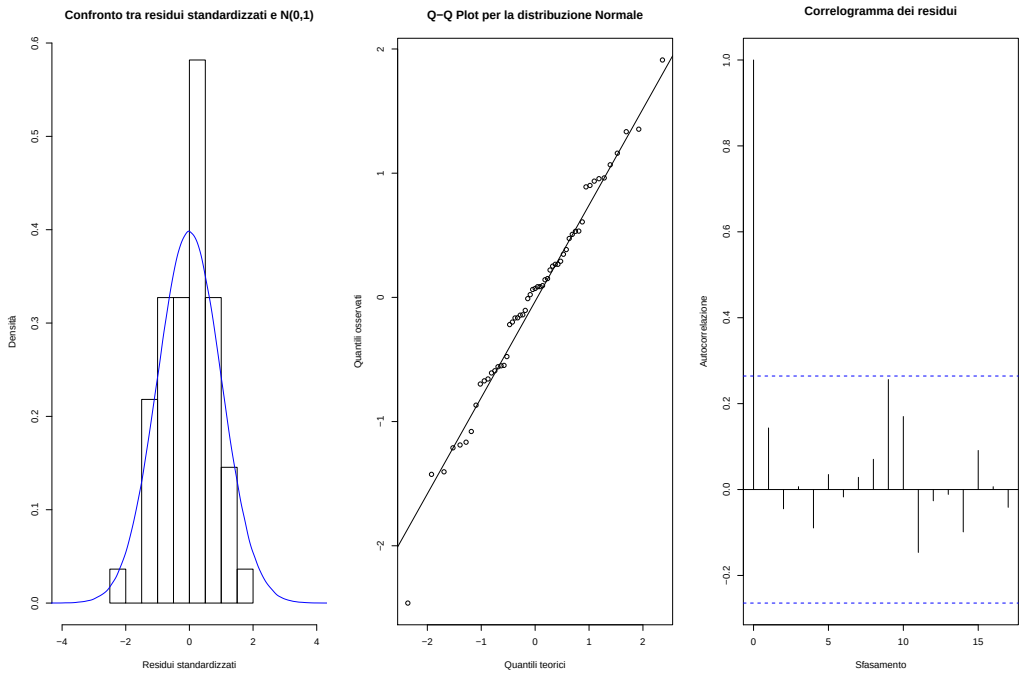


12. Marche: anche in questo caso, il *p-value* del test di Shapiro-Wilk assume un valore elevato (0.7012), la distribuzione attorno allo zero è abbastanza regolare ma notiamo un'asimmetria verso destra: il 54.5% delle percentuali ottenute dal modello sono sottostime (figura 3.25 e tabella 3.32). D'altro canto, rileviamo un *outlier* in corrispondenza del mese di maggio 2011: la percentuale osservata è 16.07%, mentre la stima è pari a 26.06%. L'ipotesi di aleatorietà non è messa in discussione.

Tabella 3.32: *Summary* dei residui standardizzati - Marche

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.461000	-0.554700	0.071830	-0.007978	0.490800	1.911000

Figura 3.25: Analisi dei residui: normalità e autocorrelazioni - Marche

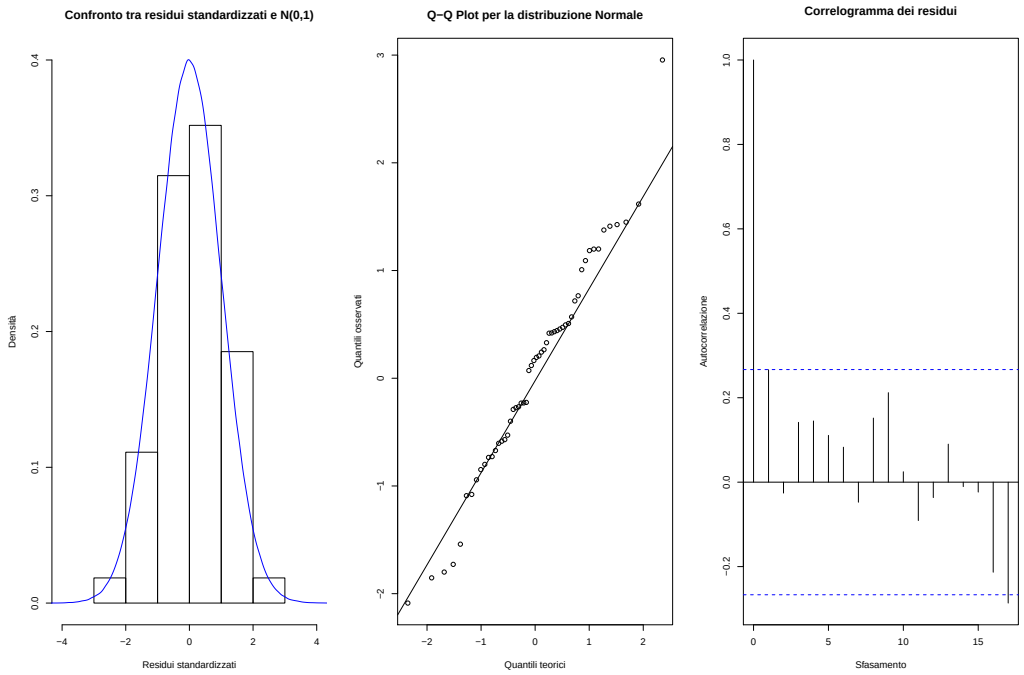


13. Lazio: considerazioni molto simili possono essere fatte per i residui calcolati sui valori relativi alla regione Lazio, poichè il *p-value* del test di Shapiro-Wilk è pari allo 0.586 ma si nota un'asimmetria verso destra che sta ad indicare una percentuale di sottostime pari al 54.5% (figura 3.26 e tabella 3.33). Una di queste sottostime è particolarmente rilevante, infatti dal boxplot dei residui viene evidenziato un *outlier*: la frazione di fumatori osservata nell'ottobre del 2007 è pari al 41.53%, mentre la stima per quel mese è 37.74%. Possiamo ritenere valida l'ipotesi di casualità.

Tabella 3.33: *Summary* dei residui standardizzati - Lazio

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.08700	-0.59910	0.17900	0.05765	0.55420	2.95500

Figura 3.26: Analisi dei residui: normalità e autocorrelazioni - Lazio

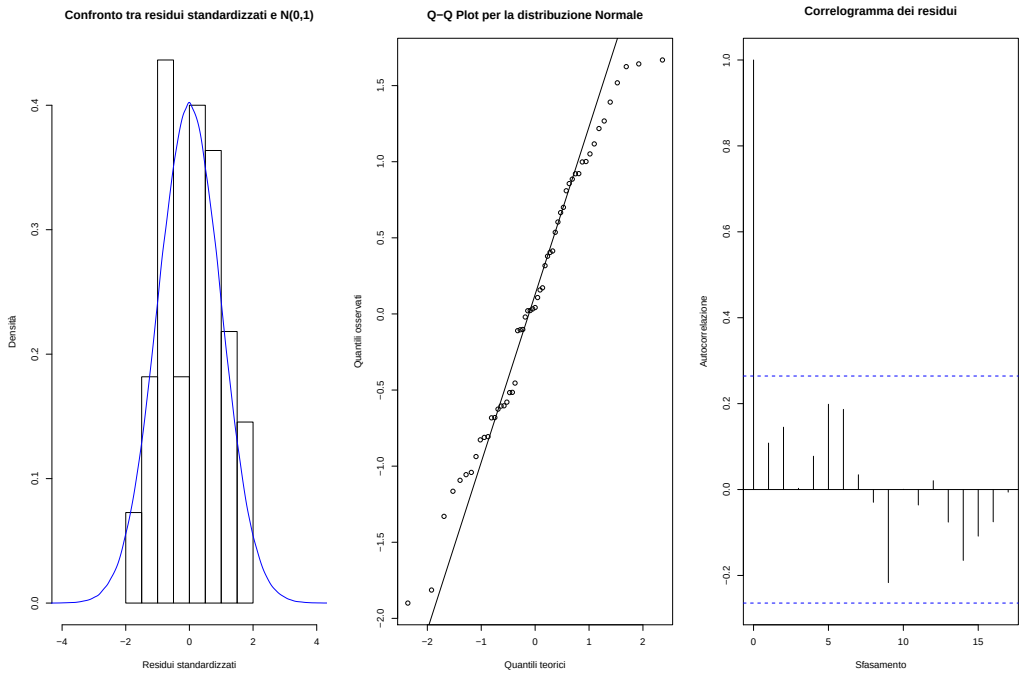


14. Abruzzo: i residui in questo caso sono piuttosto concentrati intorno allo zero, rispetto alle situazioni analizzate in precedenza, ed il *p-value* del test di Shapiro-Wilk (0.2735) è tale da consentire l'accettazione dell'ipotesi di normalità; tuttavia rileviamo un'asimmetria verso destra: più del 56% dei valori è sottostimato (figura 3.27 e tabella 3.34). Dal correlogramma non risulta significativo alcun coefficiente.

Tabella 3.34: *Summary* dei residui standardizzati - Abruzzo

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.90000	-0.61530	0.04199	0.09255	0.87090	1.66800

Figura 3.27: Analisi dei residui: normalità e autocorrelazioni - Abruzzo

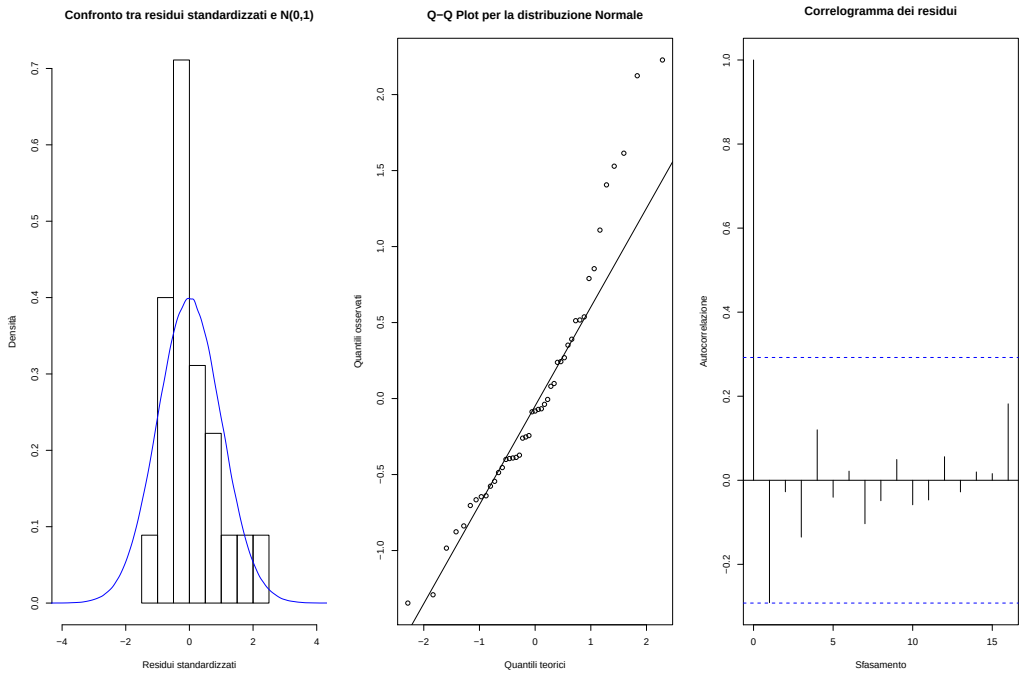


15. Molise: l'ipotesi di casualità in questo caso è confermata, mentre quella di normalità va rifiutata: il *p-value* del test di Shapiro-Wilk è 0.01395 e c'è un'evidente asimmetria verso sinistra, visto che quasi il 68% delle stime supera i dati reali, come si evince osservando il grafico 3.28 ed i valori dei residui, riassunti nella tabella 3.35. Due inoltre sono gli *outlier* riscontrati (sottostime) relativi a febbraio 2008 e giugno 2011: le percentuali osservati sono rispettivamente 48% e 45.16%, quelle stimate 29.21% e 27.74%. La bontà della retta di regressione offerta dal modello è dunque in discussione e questo può essere dovuto a due fattori: mancano le rilevazioni di 10 mensilità ed inoltre le numerosità dei campioni mensili sono piuttosto basse.

Tabella 3.35: *Summary* dei residui standardizzati - Molise

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.34600	-0.48690	-0.08203	0.03968	0.39050	2.22800

Figura 3.28: Analisi dei residui: normalità e autocorrelazioni - Molise

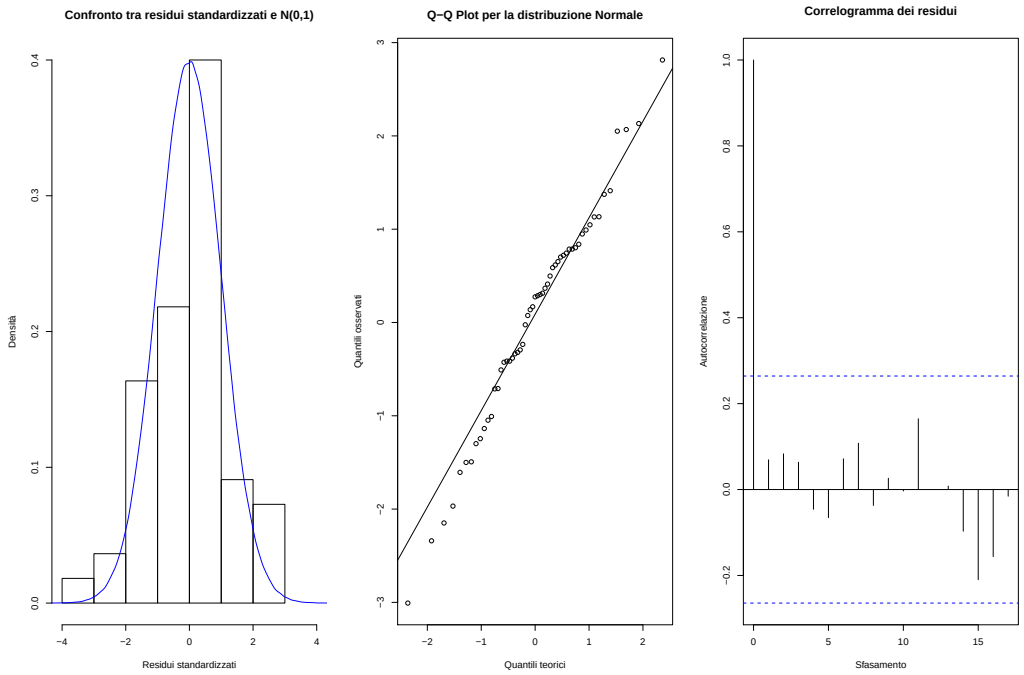


16. Campania: casualità e normalità sono confermate dal grafico delle autocorrelazioni e dal test di Shapiro-Wilk che presenta un *p-value* pari a 0.716. Tuttavia, rileviamo un'asimmetria verso destra, poichè più del 56% dei valori è sottostimato, ed un *outlier*, che si riferisce ad una sovrastima in corrispondenza di ottobre 2009, in cui è stata rilevata una percentuale pari a 22.73%, mentre il dato stimato è del 29.90% (figura 3.29 e tabella 3.36)

Tabella 3.36: *Summary* dei residui standardizzati - Campania

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.00800	-0.60810	0.27490	0.04728	0.78680	2.81400

Figura 3.29: Analisi dei residui: normalità e autocorrelazioni - Campania

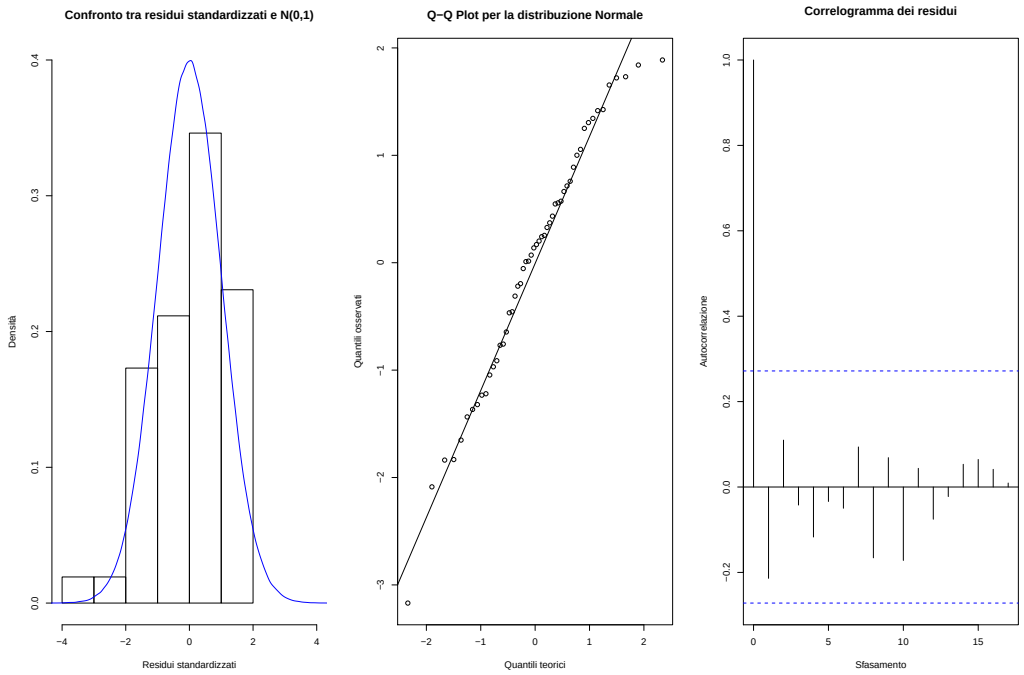


17. Puglia: il *p-value* del test di Shapiro-Wilk è tale da farci accettare l'ipotesi di normalità (0.3257). Dal grafico 3.30 e dai valori dei residui 3.37 riscontriamo un'asimmetria verso destra (il 54.5% dei valori è sottostimato), anche se i residui più consistenti in valore assoluto fanno riferimento a delle sottostime. L'analisi delle autocorrelazioni non mette in dubbio la casualità dei residui.

Tabella 3.37: *Summary* dei residui standardizzati - Puglia

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.16900	-0.80460	0.15360	0.01192	0.79120	1.88800

Figura 3.30: Analisi dei residui: normalità e autocorrelazioni - Puglia

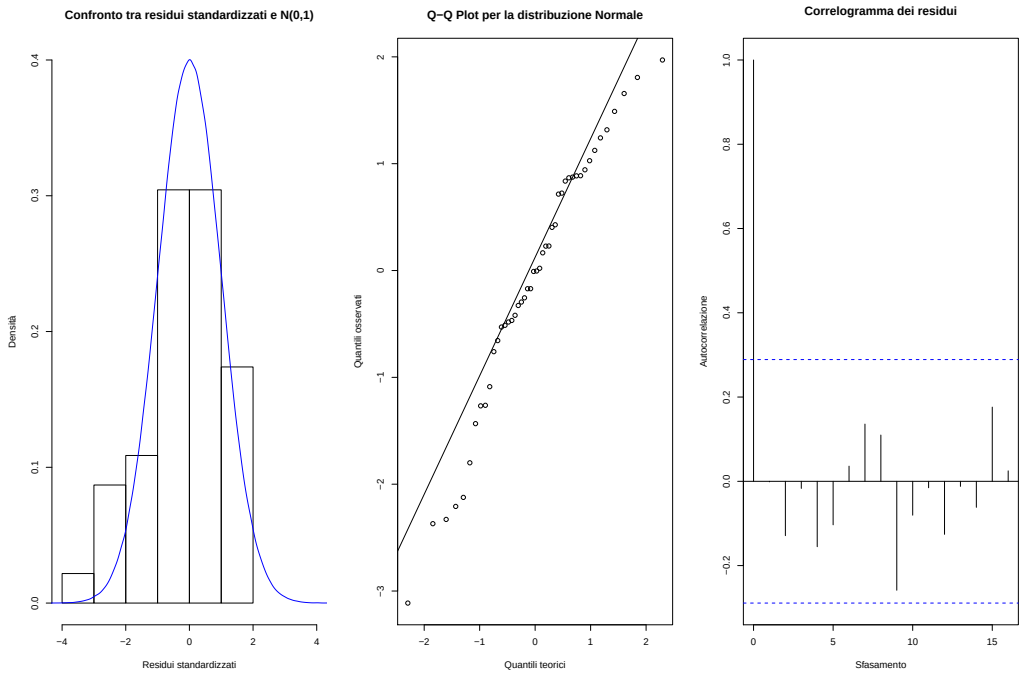


18. Basilicata: nonostante un'asimmetria verso sinistra (il 60% dei valori è sovrastimato) ed un *outlier* in corrispondenza di settembre 2008 (il dato reale è 12.24%, quello stimato è 25.40%), il test di Shapiro-Wilk presenta un *p-value* pari a 0.1911 che ci porta a non rifiutare l'ipotesi di normalità. Nessun coefficiente di autocorrelazione risulta significativamente diverso da zero. In dettaglio, figura 3.31 e tabella 3.38.

Tabella 3.38: *Summary* dei residui standardizzati - Basilicata

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.113000	-0.624100	-0.006444	-0.091320	0.872800	1.971000

Figura 3.31: Analisi dei residui: normalità e autocorrelazioni - Basilicata

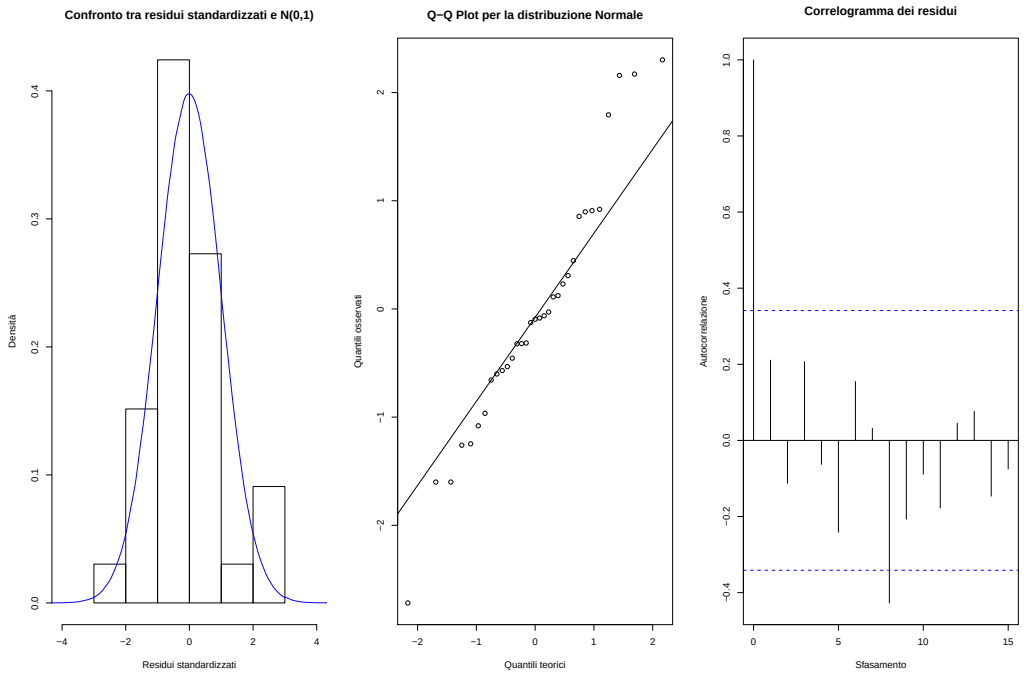


19. Calabria: per questa regione mancano le rilevazioni di più di 20 mensilità. Ci aspettiamo dunque di riscontrare una distribuzione poco omogenea dei residui ed infatti dal grafico 3.32 e dai valori riportati sinteticamente nella tabella 3.39 notiamo una forte asimmetria verso sinistra, dunque buona parte dei valori offerti dal modello (più del 76%) risulta essere una sovrastima. Inoltre, tra i residui individuiamo ben tre *outlier*: due sottostime, relative ad agosto 2009 (dato osservato pari a 37.35%, stima pari a 27%) ed a febbraio 2011 (dato osservato pari a 40.91%, stima pari a 26.08%) ed una sovrastima, corrispondente ad ottobre 2010 (dato osservato pari a 12.16%, stima pari a 25.66%). Nonostante ciò, il test di Shapiro-Wilk offre un *p-value* superiore allo 0.05, pari a 0.3316. Il grafico delle autocorrelazioni fa emergere un coefficiente significativamente diverso da zero, ma anche in questo caso possiamo fare riferimento alla spiegazione di pagina 95.

Tabella 3.39: *Summary* dei residui standardizzati - Calabria

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.71700	-0.60130	-0.09508	-0.04274	0.44640	2.30300

Figura 3.32: Analisi dei residui: normalità e autocorrelazioni - Calabria

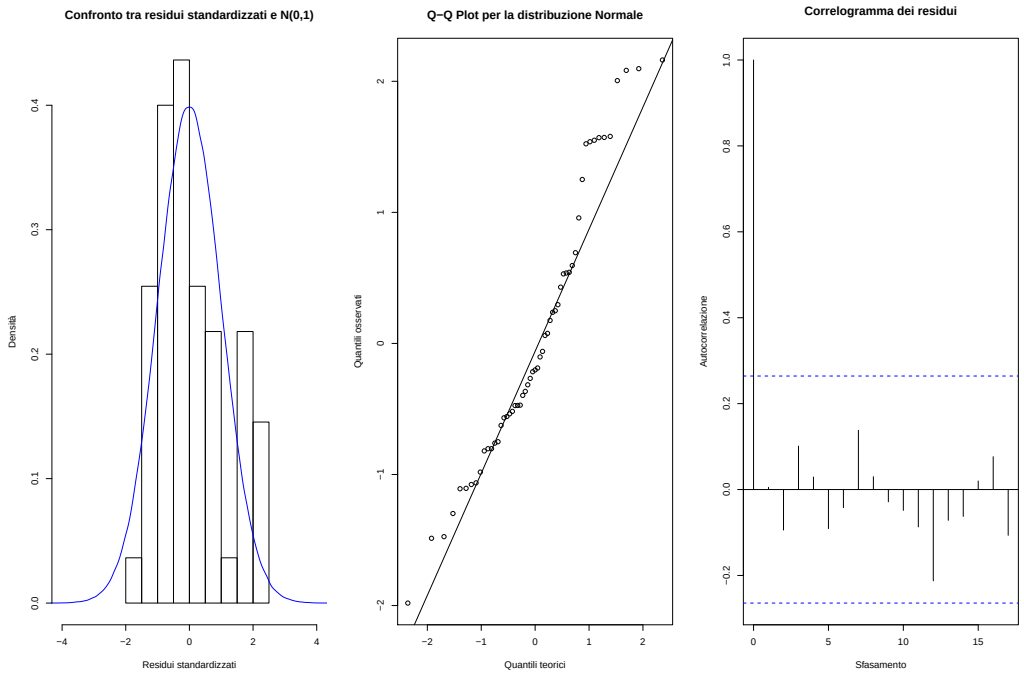


20. Sicilia: in questo caso, l'ipotesi di casualità è accettata, poichè nessun ρ_k risulta significativamente diverso da zero, mentre il test di Shapiro-Wilk, che presenta un *p-value* dello 0.02181, suggerisce di rifiutare l'ipotesi di normalità. Dal grafico 3.33 e dai valori riassunti in tabella 3.40 si nota una qualche irregolarità nella distribuzione dei residui ed un'asimmetria verso sinistra, infatti più del 56% dei dati risulta essere sovrastimato dal modello.

Tabella 3.40: *Summary* dei residui standardizzati - Sicilia

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.98100	-0.68770	-0.20210	0.04476	0.56850	2.16300

Figura 3.33: Analisi dei residui: normalità e autocorrelazioni - Sicilia

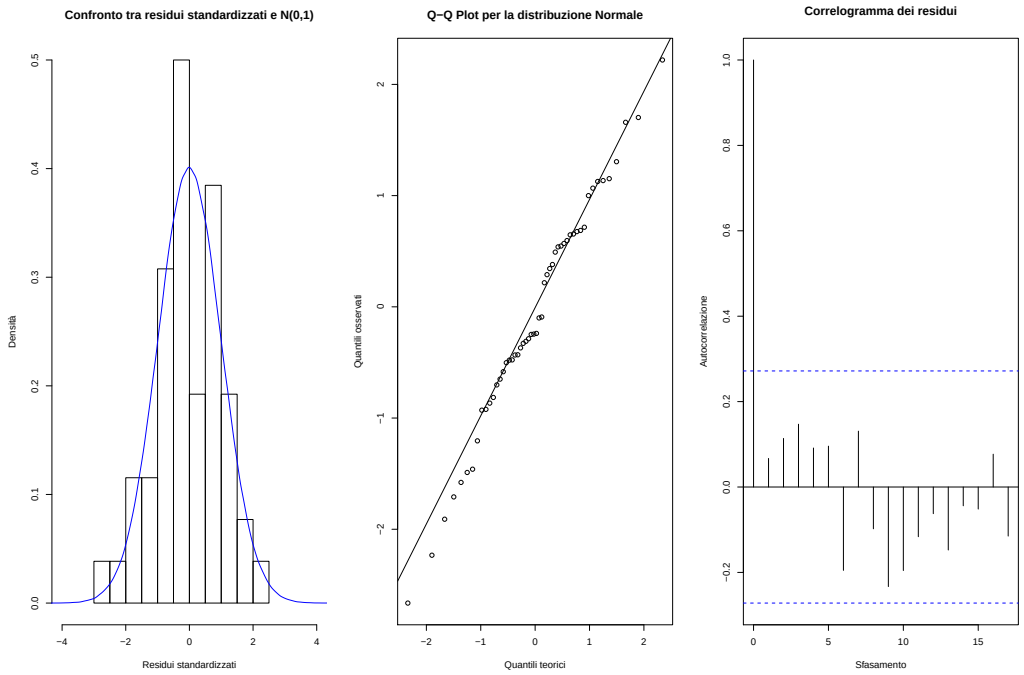


21. Sardegna: anche in questo caso siamo in presenza di un'asimmetria verso sinistra (più del 58% dei valori è sovrastimato), ma la distribuzione dei residui segue maggiormente la forma "a campana" e l'ipotesi di normalità è confermata, visto che il *p-value* del test di Shapiro-Wilk è pari allo 0.8477 (figura 3.34 e tabella 3.41). Nessun dubbio sull'aleatorietà dei residui.

Tabella 3.41: *Summary* dei residui standardizzati - Sardegna

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.66400	-0.66300	-0.24210	-0.08722	0.64910	2.21900

Figura 3.34: Analisi dei residui: normalità e autocorrelazioni - Sardegna



Proviamo a costruire dei modelli contenenti una componente di trend polinomiale di secondo e terzo grado, mi ci accorgiamo che il valore del BIC aumenta (tabella 3.42); quindi, il modello presentato in precedenza resta quello scelto per la nostra analisi.

Tabella 3.42: Confronto tra BIC - partizionamento per classe d'età

	Grado 1	Grado 2	Grado 3
BIC	1194	1210	1245

3.3.2 Modelli con due variabili di stratificazione

Sviluppiamo dei modelli che, se ipotizziamo di rappresentare una tendenza di fondo tramite un polinomio di primo grado e di considerare due variabili per determinare la partizione del campione di riferimento, assumono questa formulazione:

$$y_i = logit^{-1}(\alpha_{j[i],k[i]} + \beta_{j[i],k[i]} t_i) = \frac{e^{\alpha_{j[i],k[i]} + \beta_{j[i],k[i]} t_i}}{1 + e^{\alpha_{j[i],k[i]} + \beta_{j[i],k[i]} t_i}},$$

dove

$$\begin{pmatrix}\alpha_{jk}\\ \beta_{jk}\end{pmatrix} = \begin{pmatrix}\mu_\alpha\\ \mu_\beta\end{pmatrix} + \begin{pmatrix}\eta_j^{var1}\\ \zeta_j^{var1}\end{pmatrix} + \begin{pmatrix}\eta_k^{var2}\\ \zeta_k^{var2}\end{pmatrix} + \begin{pmatrix}\eta_{jk}^{var1\times var2}\\ \zeta_{jk}^{var1\times var2}\end{pmatrix}$$

con

$$\begin{pmatrix}\eta_j^{var1}\\ \zeta_j^{var1}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma^{var1}\right), \qquad j = 1, \ldots, J.$$

$$\begin{pmatrix}\eta_k^{var1}\\ \zeta_k^{var1}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma^{var2}\right), \qquad k = 1, \ldots, K.$$

$$\begin{pmatrix}\eta_{jk}^{var1\times var2}\\ \zeta_{jk}^{var1\times var2}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma^{var1\times var2}\right), \qquad j = 1, \ldots, J; k = 1, \ldots, K.$$

Il vettore formato da η_j^{var1} e ζ_j^{var1} rappresenta gli effetti (scostamenti) che la prima variabile considerata per la stratificazione determina sull'intercetta e sul coefficiente angolare della retta di regressione stimata per ognuno dei $J \times K$ gruppi; discorso equivalente per il vettore formato da η_k^{var2} e ζ_k^{var2}, relativo alla seconda variabile. Gli elementi del terzo vettore, $\eta_{jk}^{var1\times var2}$ e $\zeta_{jk}^{var1\times var2}$, rappresentano invece gli effetti determinati dall'interazione tra le due variabili. Lo scostamento complessivo di ogni gruppo, per intercetta e coefficiente, si ottiene addizionando gli elementi dei tre vettori.

3.3.2.1 Variabili di stratificazione: classe d'età & sesso

Se suddividiamo gli intervistati in base alla classe d'età a cui appartengono ed in base al loro sesso, otteniamo 6 gruppi; nella tabella 3.43 sono riportate le numerosità relative di ogni gruppo, calcolate sulla numerosità totale del campione.

Tabella 3.43: Scomposizione - partizionamento per classe d'età & sesso

	Scomposizione in percentuale
18-34, M	14.04
18-34, F	14.19
35-49, M	17.17
35-49, F	17.38
50-69, M	18.04
50-69, F	19.17

Costruiamo un modello contenente una componente di trend lineare. Nella tabella 3.44 sono contenute le stime delle medie μ_α e μ_β dei parametri variabili ed i coefficienti di correlazione tra intercetta e coefficiente angolare, ρ_{eta}, ρ_{sesso} e $\rho_{eta\&sesso}$. Questi ultimi tre valori, considerando due variabili per il partizionamento del campione di riferimento, sono difficili da interpretare senza l'appoggio di altri valori concreti; ci affidiamo dunque all'analisi degli scostamenti calcolati per ogni gruppo per dare un'interpretazione ad essi ed, in generale, al modello costruito.

Tabella 3.44: Stime dei parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età & sesso

	Stima	Std. Error	t value	*p-value*	Significatività
μ_α	-0.8850543	0.2295698	-3.855	0.000116	***
μ_β	-0.0015700	0.0006108	-2.570	0.010157	*

ρ_{eta}	ρ_{sesso}	$\rho_{eta\&sesso}$
-1	-1	1

Nelle tabelle 3.45, 3.46, 3.47 sono riportati, per ognuno dei 6 gruppi, i valori degli effetti determinati sull'intercetta e sul coefficiente angolare della retta di regressione dalle variabili che individuano la classe d'età, il sesso e dalla loro interazione. Grazie ai dati contenuti nelle tabelle siamo in grado di fare delle considerazioni su quanto stimato dal modello.

I segni dei tre tipi di scostamenti sono determinati dai coefficienti di correlazione tra intercetta e coefficiente angolare: se consideriamo gli effetti determinati dall'appartenenza ad una certa classe d'età, il segno dello scostamento riguardante il coefficiente angolare è opposto a quello relativo all'intercetta; stesso discorso vale per il sesso; gli scostamenti determinati dall'interazione tra le due variabili hanno invece segno uguale. Come spiegato in precedenza per i modelli costruiti basandosi su una sola variabile di stratificazione, nel caso di $\rho = -1$, uno scostamento positivo dell'intercetta (ossia un livello di partenza più elevato

della media) è accompagnato da uno scostamento negativo del coefficiente (ossia un incremento della rapidità di decrescita, nel caso di coefficiente medio negativo; oppure una diminuzione della rapidità di crescita, nel caso di coefficiente positivo).

Tabella 3.45: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età & sesso (variabile: classe d'età)

Gruppo	η_j^{eta}	ζ_j^{eta}
18-34	0.29211409	-0.0009170709
35-49	0.03984058	-0.0001250766
50-69	-0.33201808	0.0010423466

Tabella 3.46: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età & sesso (variabile: sesso)

Gruppo	η_j^{sesso}	ζ_j^{sesso}
M	0.2121479	-3.029377e-05
F	-0.2120597	3.028117e-05

Tabella 3.47: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età & sesso (interazione tra le due variabili)

Gruppo	$\eta_j^{eta\&sesso}$	$\zeta_j^{eta\&sesso}$
18-34 M	0.05979039	2.663226e-04
18-34 F	-0.07093257	-3.159529e-04
35-49 M	0.04401048	1.960346e-04
35-49 F	-0.00992794	-4.422173e-05
50-69 M	-0.07197428	-3.205929e-04
50-69 F	0.04906595	2.185530e-04

La tabella 3.48 riporta gli scostamenti totali per ciascun gruppo, ottenuti sommando i relativi valori contenuti nelle tabelle 3.45, 3.46 e 3.47. Il grafico 3.35

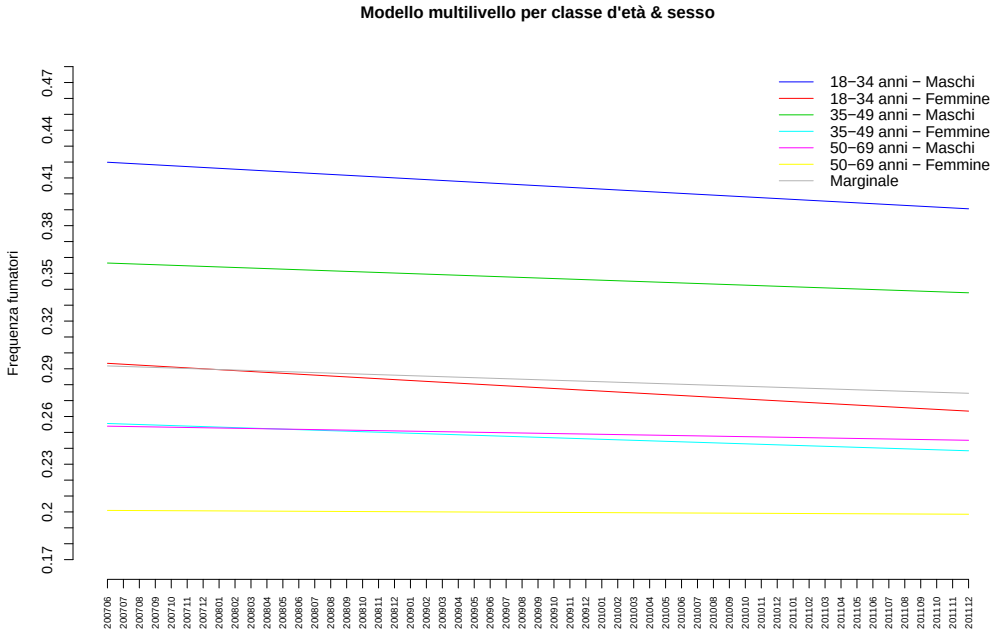
illustra le rette di regressione stimate dal modello, permettendoci di confrontarle visivamente tra loro.

La frazione di fumatori riscontrata tra i maschi aventi un'età compresa tra i 18 ed i 34 anni supera di più di 10 punti percentuali durante tutto il periodo di tempo considerato quella media e quella rilevata tra le femmine della stessa fascia d'età; il fenomeno all'interno di questa sottopopolazione presenta però una decrescita lievemente più rapida rispetto alla tendenza media. Anche gli uomini tra i 35 ed i 49 anni fumano più della media, mentre quelli appartenenti alla fascia 50-69 anni si attestano circa 2-3 punti percentuali sotto la media, così come le donne tra i 35 ed i 49 anni. Tra le donne con età compresa tra i 50 ed i 69 anni si registra una percentuale pressochè costante di fumatrici, pari al 20%, valore decisamente più basso rispetto a quello riscontrato a livello generale.

Tabella 3.48: Stime degli scostamenti totali dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età & sesso)

Gruppo	η_j	ζ_j
18-34 M	0.564052377	-6.810421e-04
18-34 F	0.009121869	-1.202743e-03
35-49 M	0.295998954	4.066421e-05
35-49 F	-0.182147011	-1.390172e-04
50-69 M	-0.191844466	6.914599e-04
50-69 F	-0.495011783	1.291181e-03

Figura 3.35: Rappresentazione del modello multilivello (con componente di trend lineare) - partizionamento per classe d'età & sesso



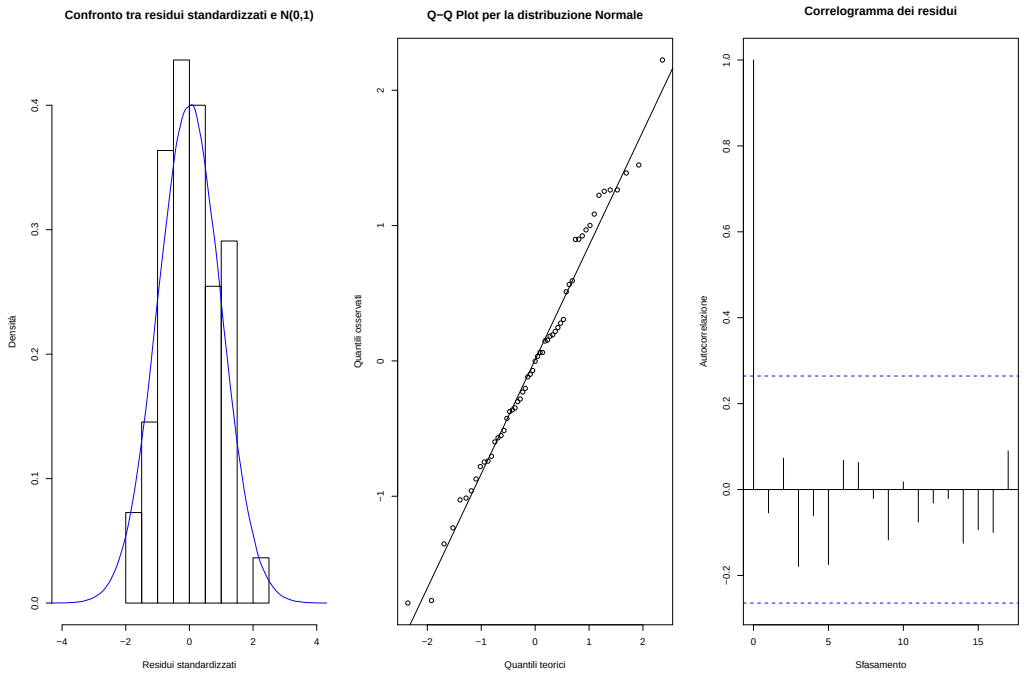
Passiamo all'analisi dei residui standardizzati.

1. 18-34 anni, maschi: i residui sembrano seguire una distribuzione Normale, infatti il *p-value* ottenuto dal test di Shapiro-Wilk è pari allo 0.7339 e non vengono riscontrate asimmetrie e nemmeno valori anomali. Il correlogramma non fa emergere alcun coefficiente significativamente diverso da zero (tabella 3.49 e figura 3.36).

   Tabella 3.49: *Summary* dei residui standardizzati - 18-34 anni, maschi

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.789000	-0.560000	-0.002292	0.024520	0.578400	2.224000

Figura 3.36: Analisi dei residui: normalità e autocorrelazioni - 18-34 anni, maschi

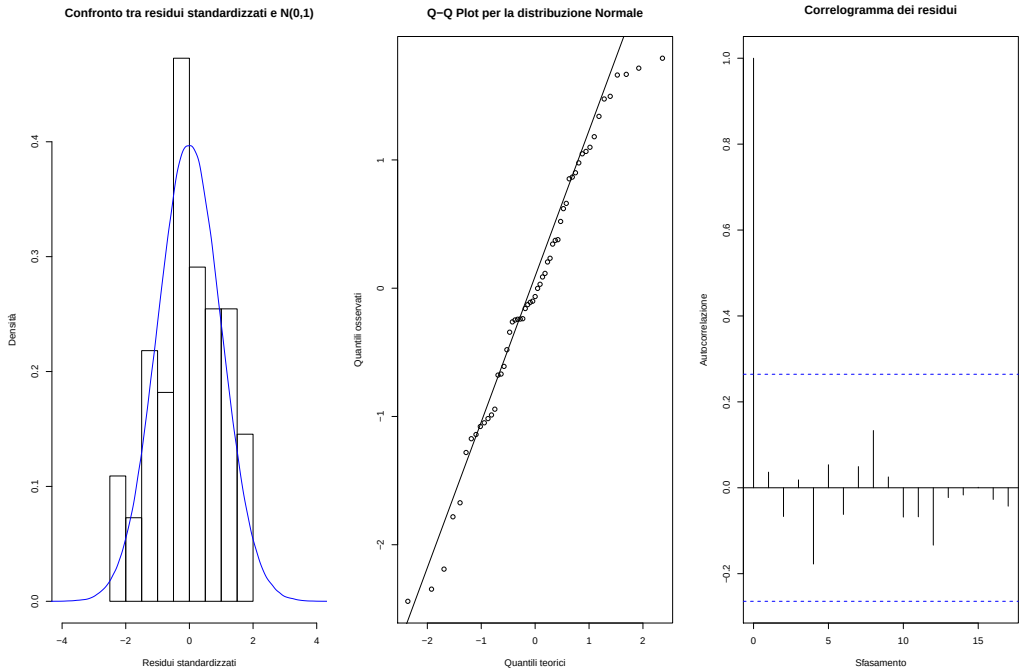


2. 18-34 anni, femmine: la distribuzione dei residui calcolati all'interno di questo gruppo è più irregolare rispetto alla precedente,tuttavia possiamo accettare l'ipotesi di normalità: il *p-value* del test di Shapiro-Wilk è inferiore a quello riscontrato in precedenza, ma comunque superiore al livello soglia (0.2909), inoltre non notiamo particolari asimmetrie, nè *outlier* (tabella 3.50 e figura 3.37). Il grafico delle autocorrelazioni conferma l'ipotesi di casualità.

Tabella 3.50: *Summary* dei residui standardizzati - 18-34 anni, femmine

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.44100	-0.67330	-0.06517	-0.01764	0.85950	1.79300

Figura 3.37: Analisi dei residui: normalità e autocorrelazioni - 18-34 anni, femmine

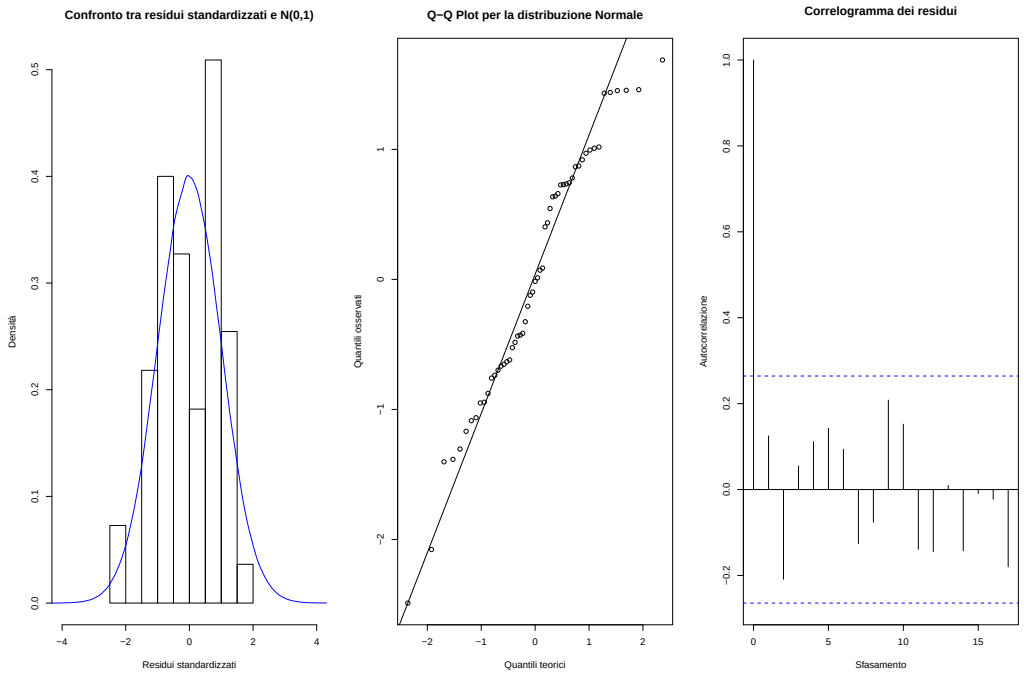


3. 35-49 anni, maschi: il *p-value* offerto dal test di Shapiro-Wilk è pari allo 0.1203: un valore non estremamente alto, ma che ci porta comunque a non rifiutare l'ipotesi di normalità. Dal grafico 3.38 notiamo che i residui non si distribuiscono in modo regolare secondo una forma "a campana", però non vi è alcuna asimmetria marcata e non vengono rilevati valori anomali (tabella 3.51). L'aleatorietà è confermata dal correlogramma.

Tabella 3.51: *Summary* dei residui standardizzati - 35-49 anni, maschi

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.489000	-0.682800	-0.015050	0.003707	0.761300	1.687000

Figura 3.38: Analisi dei residui: normalità e autocorrelazioni - 35-49 anni, maschi

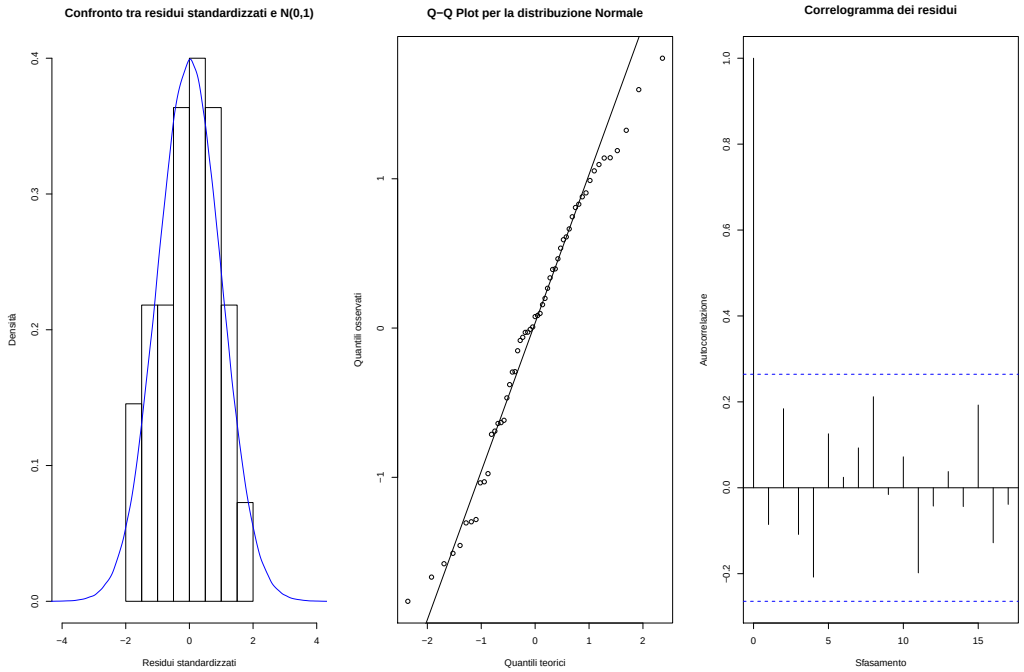


4. 35-49 anni, femmine: in questo caso la distribuzione dei residui attorno allo zero è più regolare, tuttavia viene riscontrata una leggera asimmetria verso destra: quasi il 53% dei valori è sottostimato. Il *p-value* del test di Shapiro-Wilk è dello 0.3214, tale da farci accettare l'ipotesi di normalità (tabella 3.52 e figura 3.39). Non risulta significativamente diverso da zero alcun coefficiente di autocorrelazione.

Tabella 3.52: *Summary* dei residui standardizzati - 35-49 anni, femmine

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.831000	-0.636400	0.076690	0.005801	0.704700	1.808000

Figura 3.39: Analisi dei residui: normalità e autocorrelazioni - 35-49 anni, femmine

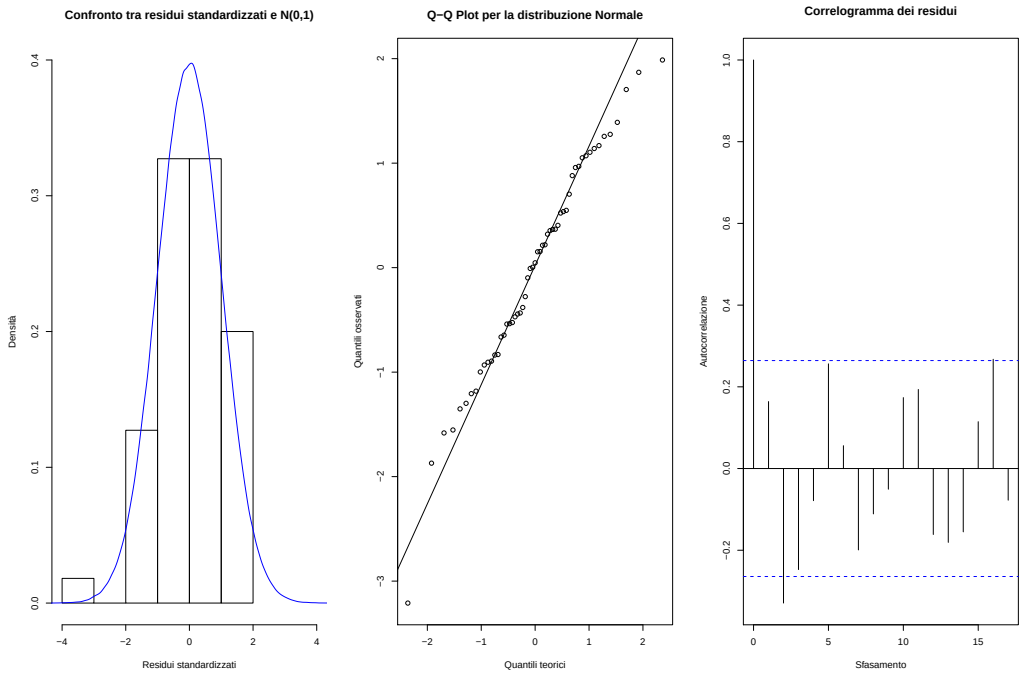


5. 50-69 anni, maschi: come nel caso precedente, i residui attorno allo zero seguono una distribuzione "a campana" ed il *p-value* del test di Shapiro-Wilk è tale da indicarci l'accettazione dell'ipotesi di normalità (0.5229). Rileviamo una lieve asimmetria verso destra ed un *outlier*, relativo ad una sovrastima: la percentuale di fumatori rilevata nel mese di maggio 2011 è pari al 19.88%, mentre quella stimata è 24.62% (tabella 3.53 e figura 3.40). Osservando il correlogramma, ci accorgiamo della significatività di un paio di coefficienti, ma accettiamo l'ipotesi di aleatorietà, poichè ci rifacciamo alla spiegazione esposta a pagina 95.

Tabella 3.53: *Summary* dei residui standardizzati - 50-69 anni, maschi

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.21000	-0.74770	0.04631	-0.01706	0.79190	1.98700

Figura 3.40: Analisi dei residui: normalità e autocorrelazioni - 50-69 anni, maschi

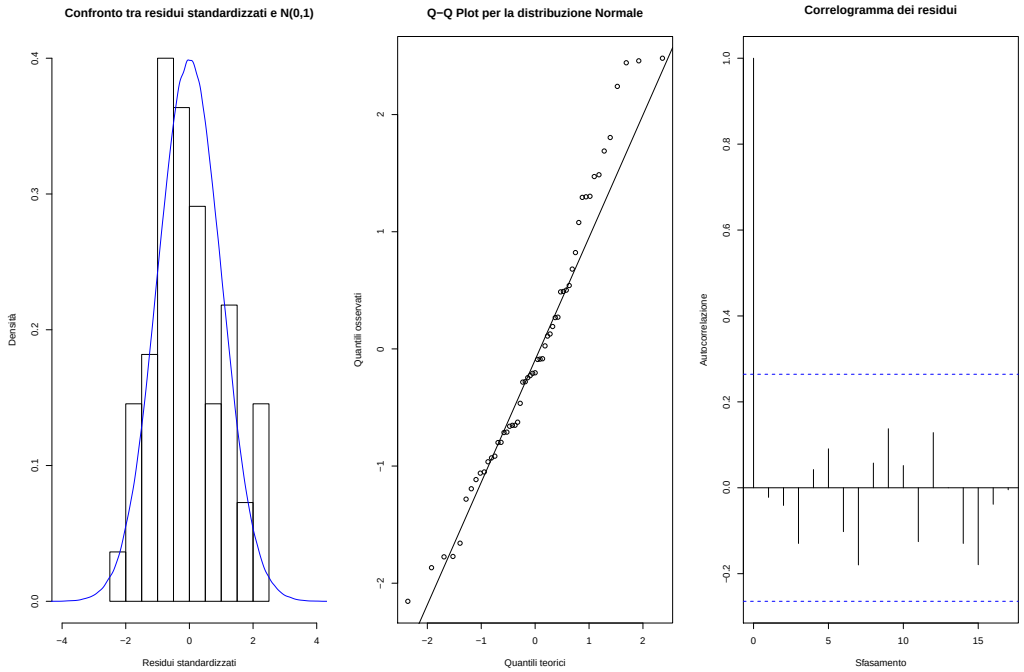


6. 50-69 anni, femmine: i residui calcolati all'interno di questo gruppo si distribuiscono asimmetricamente, infatti quasi il 57% dei dati offerti dal modello risultano essere delle sovrastime. Nonostante ciò, il *p-value* del test di Shapiro-Wilk è superiore alla soglia, anche se non di molto (0.1135); l'ipotesi di normalità viene dunque accettata, così come quella di casualità (tabella 3.54 e figura 3.41).

Tabella 3.54: *Summary* dei residui standardizzati - 50-69 anni, femmine

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.21000	-0.74770	0.04631	-0.01706	0.79190	1.98700

Figura 3.41: Analisi dei residui: normalità e autocorrelazioni - 50-69 anni, femmine



Confrontiamo i BIC risultanti dai modelli con componente di trend polinomiale di primo, secondo e terzo grado (tabella 3.55): come in tutti i casi analizzati in precedenza, il modello preferibile è quello che presenta un trend lineare.

Tabella 3.55: Confronto tra BIC - partizionamento per classe d'età & sesso

	Grado 1	Grado 2	Grado 3
BIC	361.9	534.6	582.6

3.3.2.2 Variabili di stratificazione: livello d'istruzione & sesso

Partizioniamo il campione di riferimento in base al livello d'istruzione degli intervistati ed al loro sesso ed otteniamo otto gruppi, le cui numerosità relative sono riportate nella tabella 3.56.

Modelliamo una componente di trend lineare e le stime delle medie dei parametri variabili sono contenute nella tabella 3.44, così come ρ_{istr}, ρ_{sesso}, $\rho_{istr\&sesso}$,

Tabella 3.56: Scomposizione - partizionamento per livello d'istruzione & sesso

	Scomposizione in percentuale
Nessuno/Elementare, M	4.81
Nessuno/Elementare, F	7.10
Scuola media inferiore, M	16.66
Scuola media inferiore, F	14.65
Scuola media superiore, M	21.86
Scuola media superiore, F	21.85
Laurea/Diploma universitario, M	5.92
Laurea/Diploma universitario, F	7.15

che misurano la correlazione che intercorre tra l'intercetta ed il coefficiente angolare, relativamente ad ognuna delle variabili prese in considerazione per il partizionamento, ed alla loro interazione. In questo caso, i tre coefficienti sono tutti pari a 1: un livello di partenza più alto della media è accompagnato da un proporzionale calo della rapidità di decrescita del fenomeno.

Tabella 3.57: Stime dei parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione & sesso

	Stima	Std. Error	t value	*p-value*	Significatività
μ_α	-0.9934703	0.1907499	-5.208	1.91e-07	***
μ_β	-0.0019386	0.0003584	-5.408	6.36e-08	***

ρ_{istr}	ρ_{sesso}	$\rho_{istr\&sesso}$
1	1	1

Le tre componenti degli scostamenti, relative al livello d'istruzione, al sesso ed all'interazione tra le due variabili, sono riportate nelle tabelle 3.58, 3.59 e 3.60: facciamo attenzione soprattutto ai segni.

Gli scostamenti totali che caratterizzano ciascun gruppo sono riportati nella tabella 3.61 e sono stati ottenuti sommando opportunamente i vari η_j^* e ζ_j^*. Analizzando essi ed osservando attentamente il grafico (figura 3.42) risultante dalla stima delle rette di regressione, siamo in grado di formulare qualche considerazione.

Tabella 3.58: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione & sesso (variabile: livello d'istruzione)

Gruppo	η_j^{istr}	ζ_j^{istr}
Nessuno/Elementare	-0.19631814	-2.229437e-05
Scuola media inferiore	0.24230768	2.751705e-05
Scuola media superiore	0.09170059	1.041374e-05
Laurea/Diploma universitario	-0.13760458	-1.562671e-05

Tabella 3.59: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione & sesso (variabile: livello d'istruzione)

Gruppo	η_j^{sesso}	ζ_j^{sesso}
M	0.1767783	0.0001617161
F	-0.1766758	-0.0001616224

Tabella 3.60: Stime degli scostamenti dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione & sesso (interazione tra le due variabili)

Gruppo	$\eta_j^{istr\&sesso}$	$\zeta_j^{istr\&sesso}$
Nessuno/Elementare M	0.09310742	6.091298e-05
Nessuno/Elementare F	-0.20118519	-1.316199e-04
Scuola media inferiore M	0.14954790	9.783761e-05
Scuola media inferiore F	-0.01163971	-7.614961e-06
Scuola media superiore M	0.02369479	1.550167e-05
Scuola media superiore F	0.02794917	1.828498e-05
Laurea/Diploma universitario M	-0.15726260	-1.028847e-04
Laurea/Diploma universitario F	0.07584747	4.962113e-05

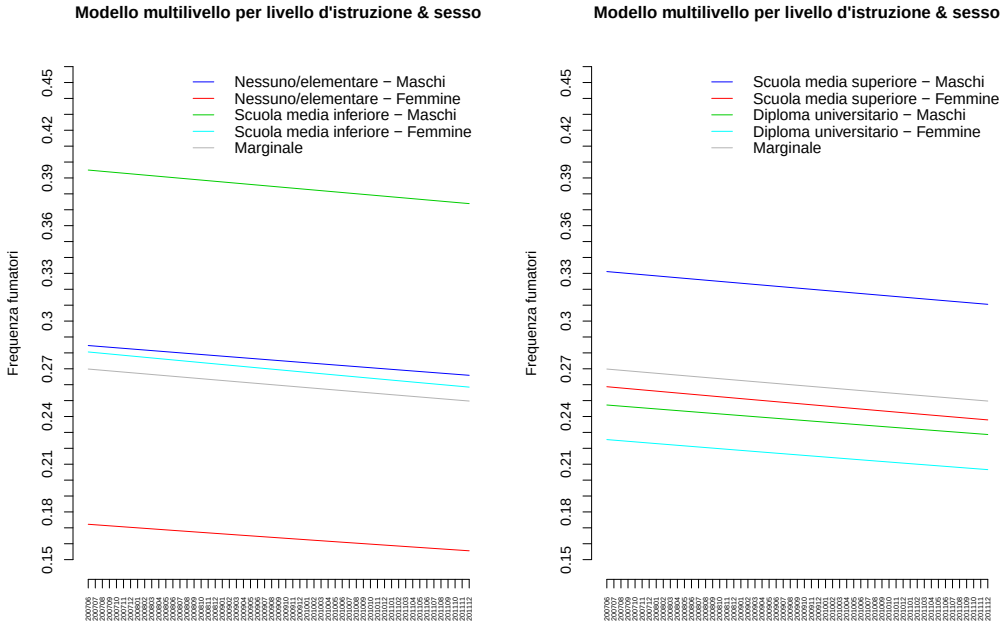
La diminuzione della percentuale di fumatori segue un ritmo all'incirca uguale in tutte le sottopopolazioni: da giugno 2007 a dicembre 2011 si registra all'interno di ognuna di esse un calo di circa 2 punti percentuali. Il gruppo all'interno del quale si registra la più alta percentuale di fumatori è quello composto da individui di sesso maschile che hanno coseguito la licenza di scuola media inferiore, seguito a ruota da quello formato dai maschi che possiedono il diploma di scuola

superiore. Poco al di sopra della media si collocano le rette relative ai maschi con nessun titolo di studio/licenza elementare ed alle femmine con la licenza di scuola media inferiore. Percentuali più basse della media si rilevano invece tra le donne che hanno conseguito il diploma di scuola superiore e tra coloro che possiedono una laurea. Infine, tra le donne con nessun titolo di studio/licenza elementare, la percentuale di fumatrici è più bassa rispetto alla media di circa 10 punti percentuali.

Tabella 3.61: Stime degli scostamenti totali dai parametri "fissi" del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione & sesso

Gruppo	η_j	ζ_j
Nessuno/Elementare M	0.07356761	2.003348e-04
Nessuno/Elementare F	-0.57417915	-3.155366e-04
Scuola media inferiore M	0.56863391	2.870708e-04
Scuola media inferiore F	0.05399215	-1.417203e-04
Scuola media superiore M	0.29217371	1.876316e-04
Scuola media superiore F	-0.05702606	-1.329237e-04
Laurea/Diploma universitario M	-0.11808885	4.320469e-05
Laurea/Diploma universitario F	-0.23843293	-1.276280e-04

Figura 3.42: Rappresentazione del modello multilivello (con componente di trend lineare) - partizionamento per livello d'istruzione & sesso



Concentriamoci sull'analisi dei residui, per avere una valutazione sulla bontà delle rette di regressione stimate.

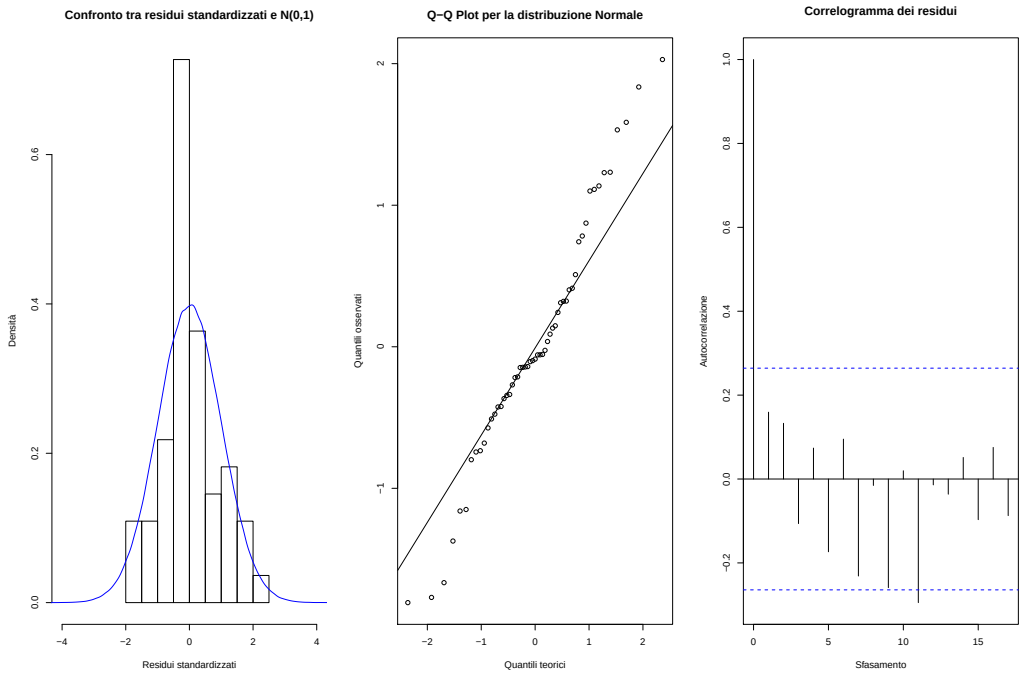
1. Nessun titolo di studio/licenza elementare, maschi: il *p-value* del test di Shapiro-Wilk è pari a 0.2536 e suggerisce l'accettazione dell'ipotesi di normalità, sebbene il grafico 3.43 ed i valori dei residui (riassunti nella tabella 3.62) denotino irregolarità nella distribuzione ed un'asimmetria verso sinistra, visto che più del 58% delle percentuali risultanti dal modello sono sovrastime. Inoltre, vi sono ben 4 valori anomali, due sovrastime e due sottostime: le prime corrispondono ai mesi di novembre 2007 (valore reale pari a 22.35%, stima pari a 28.29%) e giugno 2010 (valore reale pari a 21.43%, stima pari a 27.21%), le seconde fanno riferimento a settembre 2010 (valore reale pari a 34.27%, stima pari a 27.10%) e gennaio 2011 (valore reale pari a 34.21%, stima pari a 26.97%). L'ipotesi di casualità è confermata, nono-

stante la significatività di un coefficiente di autocorrelazione (spiegazione a pagine 95.

Tabella 3.62: *Summary* dei residui standardizzati - nessun titolo di studio/licenza elementare, maschi

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.80800	-0.42380	-0.08716	0.01863	0.40770	2.02900

Figura 3.43: Analisi dei residui: normalità e autocorrelazioni - nessun titolo di studio/licenza elementare, maschi

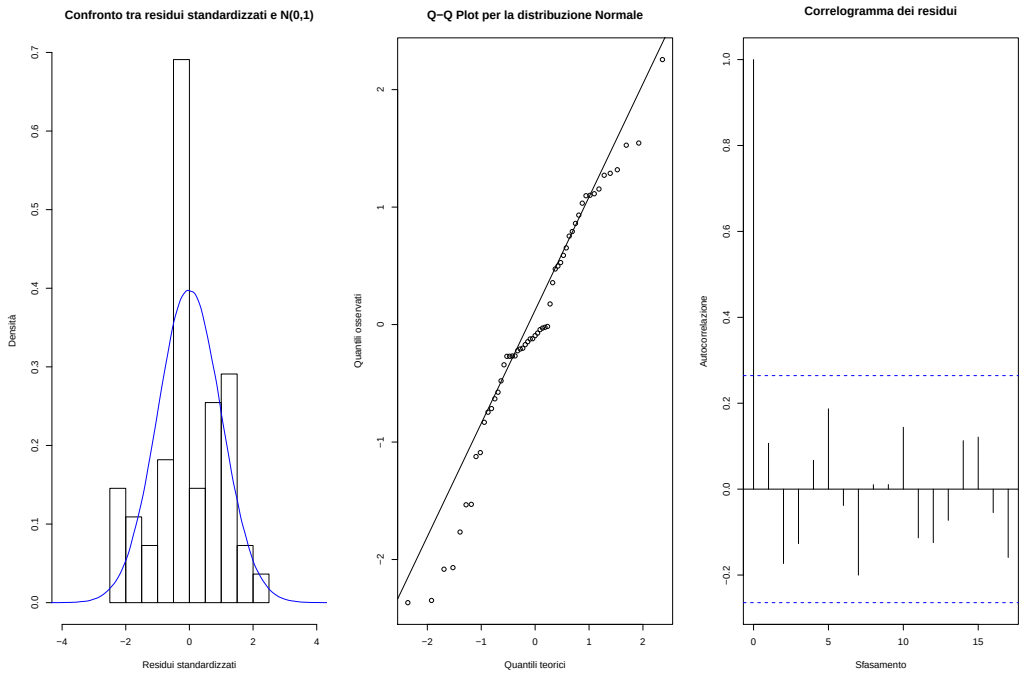


2. Nessun titolo di studio/licenza elementare, femmine: anche l'istogramma relativo a questi residui illustra una distribuzione alquanto irregolare e siamo in presenza di un'asimmetria verso sinistra: il 60% dei valori è sovrastimato (figura 3.44 e tabella 3.63). Nonostante ciò, il test di Shapiro-Wilk ci porta al non rifiuto dell'ipotesi di normalità, poichè presenta un *p-value* pari allo 0.1525. Nessun coefficiente di autocorrelazione risulta essere significativamente diverso da zero.

Tabella 3.63: *Summary* dei residui standardizzati - nessun titolo di studio/licenza elementare, femmine

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.36700	-0.52740	-0.09404	-0.02654	0.77270	2.25600

Figura 3.44: Analisi dei residui: normalità e autocorrelazioni - nessun titolo di studio/licenza elementare, femmine

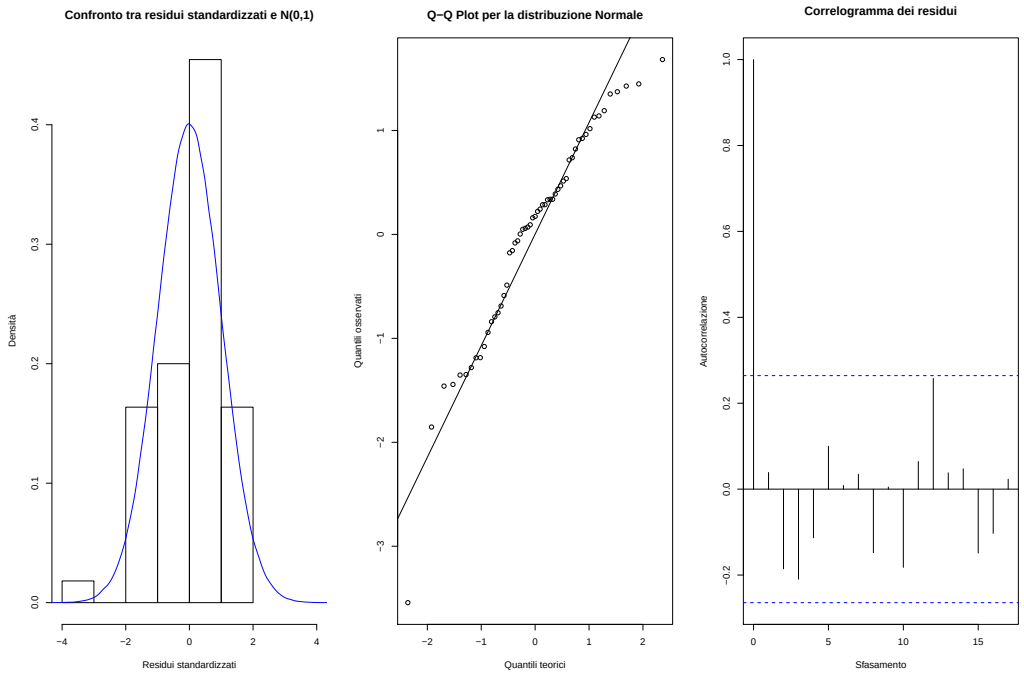


3. Licenza di scuola media inferiore, maschi: mentre l'ipotesi di aleatorietà dei residui è da accettare, quella di normalità va rifiutata. l *p-value* del test di Shapiro-Wilk è pari allo 0.01522 e si rileva un'asimmetria verso destra: più del 61% dei valori è sottostimato (figura 3.45 e tabella 3.64). Inoltre, riscontra la presenza di un *outlier* tra i residui: la percentuale osservata di fumatori relativi a maggio 2011 è pari a 31%, quella stimata è 37.66%.

Tabella 3.64: *Summary* dei residui standardizzati - licenza di scuola media inferiore, maschi

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.543000	-0.721100	0.173000	0.009404	0.727000	1.682000

Figura 3.45: Analisi dei residui: normalità e autocorrelazioni - licenza di scuola media inferiore, maschi

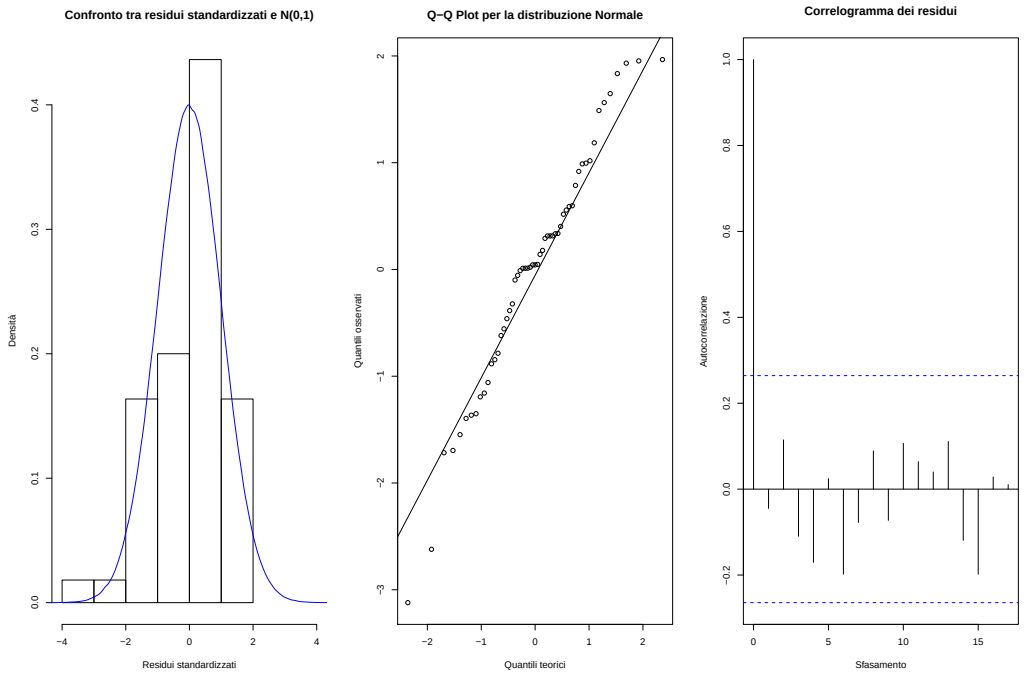


4. Licenza di scuola media inferiore, femmine: in questo caso entrambi le ipotesi sono confermate, infatti non c'è alcun coefficiente di autocorrelazione significativamente diverso da zero ed *p-value* offerto dal test di Shapiro-Wilk è dello 0.2428. Si riscontra però un'asimmetria verso destra nella distribuzione dei residui: il 60% dei valori è sottostimato. Inoltre, come nel caso precedente, ci imbattiamo in *outlier*, relativo ad una sovrastima anomala del dato di settmebre 2009: la percentuale osservata è 21.07%, quella stimata è 26.94%.

Tabella 3.65: *Summary* dei residui standardizzati - licenza di scuola media inferiore, femmine

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.121000	-0.701700	0.043400	0.002147	0.593700	1.966000

Figura 3.46: Analisi dei residui: normalità e autocorrelazioni - licenza di scuola media inferiore, femmine

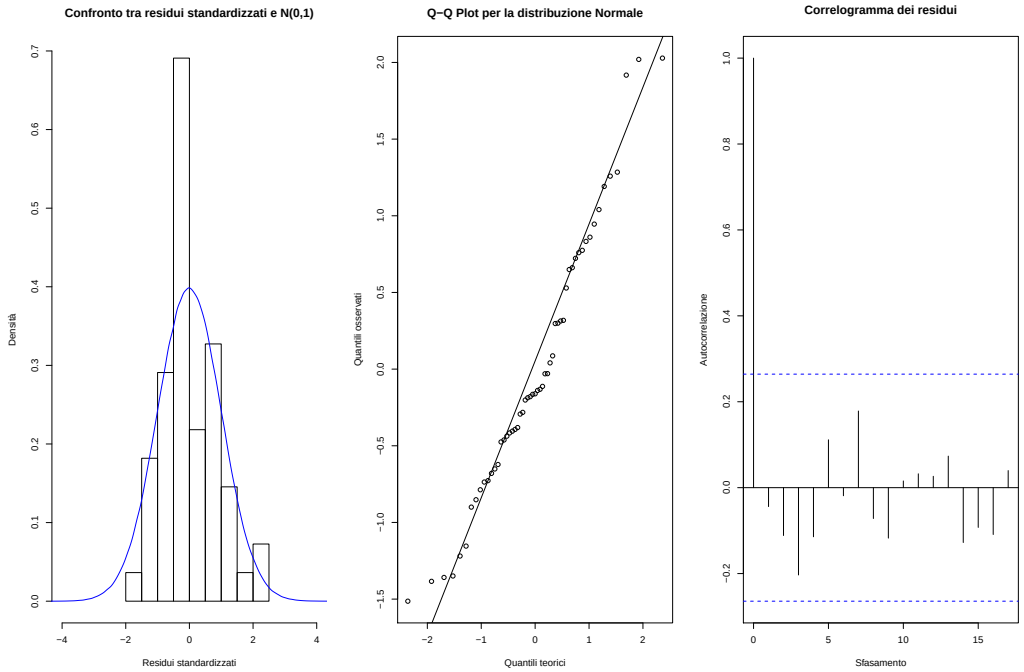


5. Diploma di scuola media superiore, maschi: distribuzione irregolare, anche in questo caso, caratterizzata da un'asimmetria verso sinistra, infatti il 60% dei valori risulta sovrastimato (figura 3.47 e tabella 3.66). Il test di Shapiro-Wilk, tuttavia, presenta un *p-value* pari a 0.124, tale da farci accettare l'ipotesi di normalità. Anche quella di casualità è confermata, come risulta dal correlogramma.

Tabella 3.66: *Summary* dei residui standardizzati - diploma di scuola media superiore, maschi

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.5140000	-0.5486000	-0.1609000	0.0002753	0.6555000	2.0280000

Figura 3.47: Analisi dei residui: normalità e autocorrelazioni - diploma di scuola media superiore, maschi

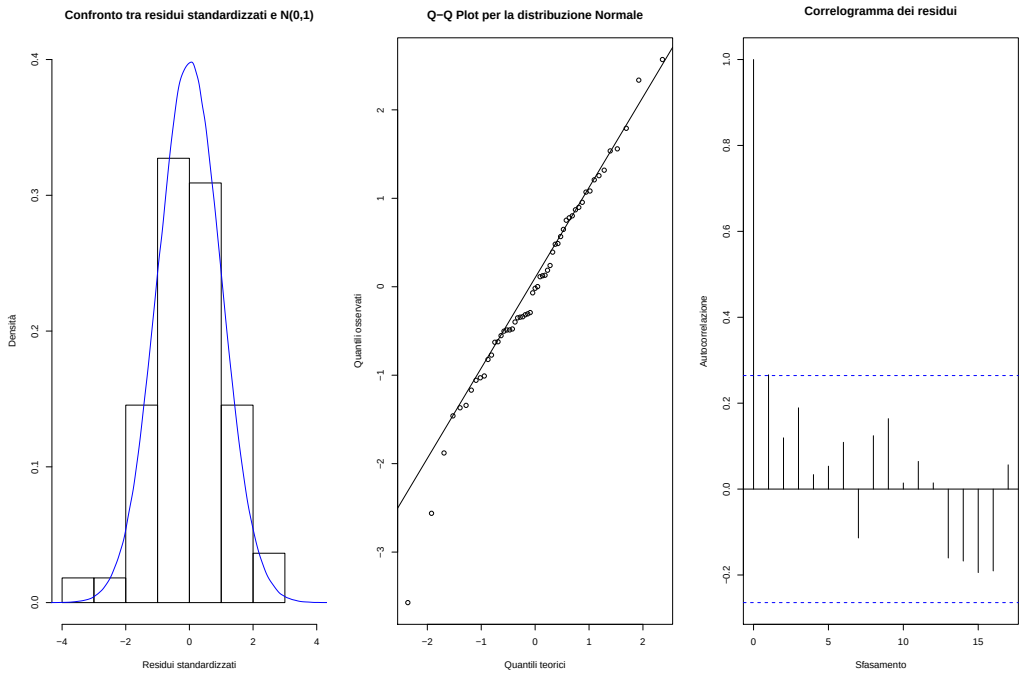


6. Diploma di scuola media superiore, femmine: l'ipotesi di normalità è confermata dal valore del *p-value* offerto dal test di Shapiro-Wilk (0.5323); non si riscontrano asimmetrie particolari. Rileviamo invece un *outlier*, relativo al mese di novembre 2009: la percentuale osservata è 19.38, quella stimata è 24.74%. L'ipotesi di aleatorietà viene accettata (figura 3.48 e tabella 3.67).

Tabella 3.67: *Summary* dei residui standardizzati - diploma di scuola media superiore, femmine

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.57100	-0.58710	-0.01735	-0.00134	0.79070	2.56900

Figura 3.48: Analisi dei residui: normalità e autocorrelazioni - diploma di scuola media superiore, femmine

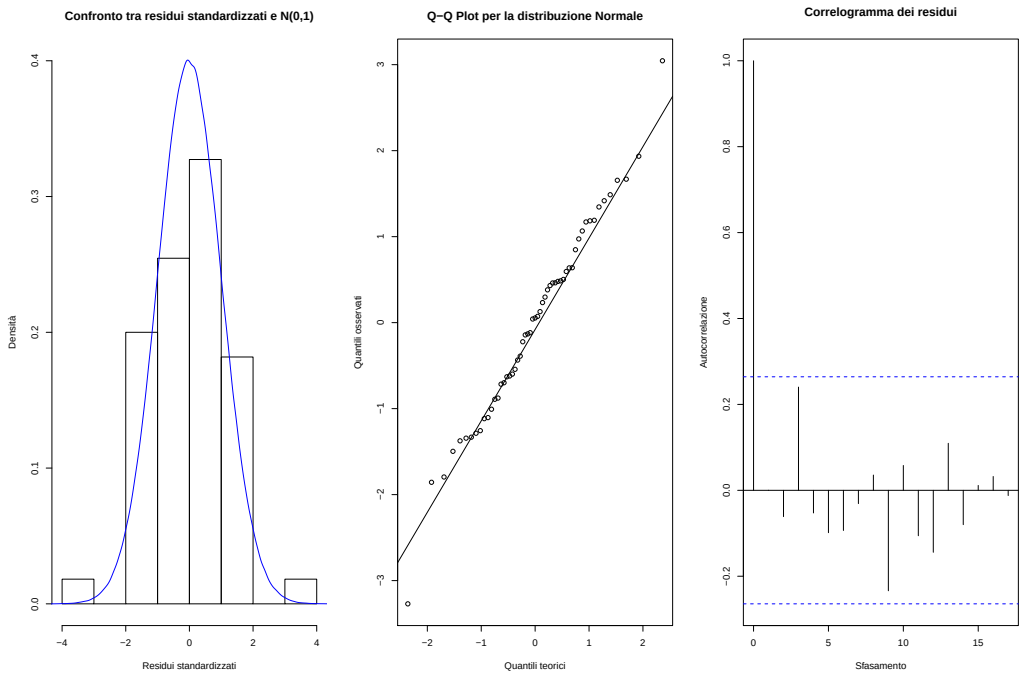


7. Laurea/Diploma universitario, maschi: come nel caso precedente, il *p-value* è particolarmente elevato (0.8897), quindi non abbiamo dubbio sull'accettazione dell'ipotesi di normalità. Non rileviamo asimmetrie anomale, mentre constatiamo che ci sono due *outlier*: il dato relativo ad ottobre 2010 è sottostimato, poichè la percentuale osservata è pari a 33.69, quella stimata è 24.58%; il dato relativo a marzo 2008 è invece sovrastimato, poichè la percentuale osservata è pari a 14.59%, quella stimata è 24.41% (figura 3.49 e tabella 3.68). Nessun dubbio nemmeno sulla casualità dei residui.

Tabella 3.68: *Summary* dei residui standardizzati - laurea/Diploma universitario, maschi

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.270000	-0.796400	0.054490	-0.006541	0.637800	3.045000

Figura 3.49: Analisi dei residui: normalità e autocorrelazioni - laurea/Diploma universitario, maschi

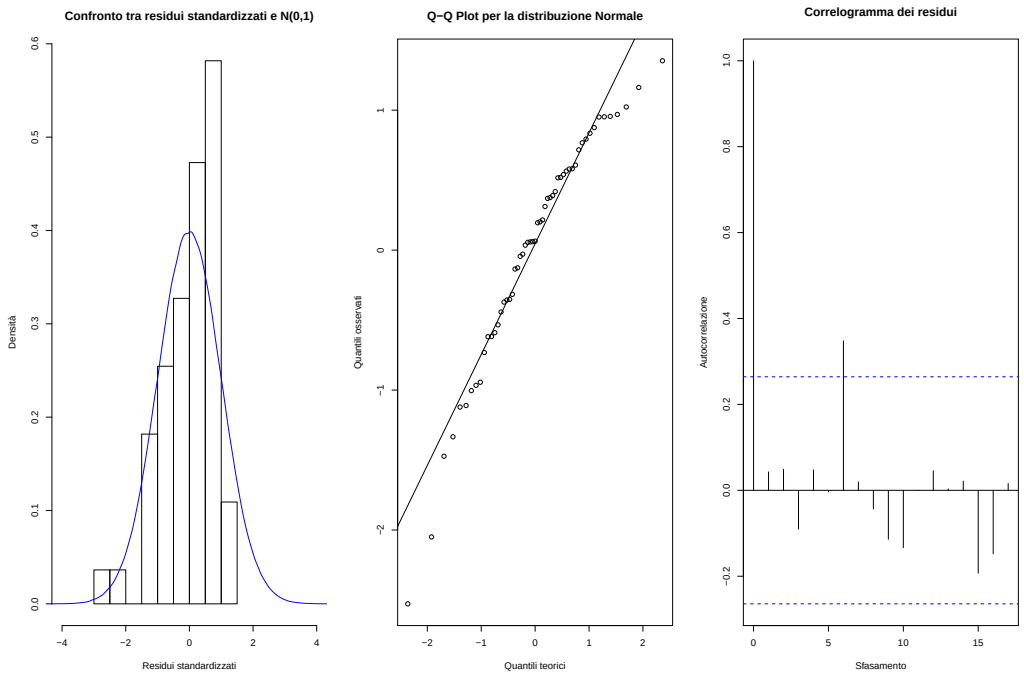


8. Laurea/Diploma universitario, femmine: l'ipotesi di normalità è da scartare, poichè il *p-value* del test di Shapiro-Wilk è pari a 0.2248. Si nota un'asimmetria verso destra: più del 58% dei valori è sottostimato; inoltre, vi è una sovrastima anomala del dato relativo a dicembre 2011: la percentuale rilevata è 14.86%, quella stimata è 21.48% (figura 3.50 e tabella 3.69). Accettiamo l'ipotesi di casualità, rifacendoci alla spiegazione di pagina 95.

Tabella 3.69: *Summary* dei residui standardizzati - laurea/diploma universitario, femmine

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.528000	-0.488100	0.063670	0.003969	0.580300	1.353000

Figura 3.50: Analisi dei residui: normalità e autocorrelazioni - laurea/diploma universitario, femmine



Confrontiamo i BIC dei modelli con componente di trend polinomiale di primo, secondo e terzo grado (tabella 3.70); il valore più basso, preferibile, lo riscontriamo nel modello con trend lineare.

Tabella 3.70: Confronto tra BIC - partizionamento per livello d'istruzione & sesso

	Grado 1	Grado 2	Grado 3
BIC	521.5	725.6	815.8

Capitolo 4

Conclusioni

Nell'elaborato sono stati presentati, dopo una prima parte descrittiva, due approcci modellistici differenti, implementati al fine di analizzare dati provenienti da serie storiche: i modelli lineari generalizzati ed i modelli multilivello.

La variabile d'interesse rappresenta la percentuale di fumatori riscontrata all'interno del campione di riferimento. Nel caso dei GLM ci siamo affidati alla distribuzione Beta-Binomiale, che ci ha consentito di modellare anche la sovra(sotto)-dispersione dei dati; mancando un'equivalente applicazione in R per quanto riguarda i modelli multilivello, per implementare questi ultimi abbiamo scelto la regressione logistica, prendendo come riferimento la distribuzione Binomiale.

Ci eravamo posti un duplice obiettivo:

1. individuare eventuali componenti di trend e stagionalità all'interno delle serie storiche, così da rilevare tendenze di fondo e comportamenti ricorrenti del fenomeno;

2. cogliere differenze tra gli andamenti della variabile d'interesse (percentuale di fumatori) riscontrati all'interno di sottopopolazioni, determinate da alcune caratteristiche socio-demografiche.

Possiamo confermare la presenza di una componente di trend, più o meno marcata a seconda del gruppo di riferimento, mentre non abbiamo rilevato

significativi movimenti stagionali del fenomeno.

I modelli multilivello ci hanno permesso di confrontare gli andamenti della variabile d'interesse registrati all'interno di particolari gruppi di individui, consentendoci di fare considerazioni dettagliate sull'evoluzione temporale del fenomeno.

Appendice

Questionario

Di seguito è riportata la versione del 2011 del questionario somministrato.

A CURA DEL COORDINATORE	
☐ Titolare	n° estrazione
☐ Sostituto	
☐ Sostituto di sostituto	_____

Sistema di sorveglianza PASSI

Progressi delle Aziende Sanitarie per la Salute in Italia

Regione _____ *ASL* _____.

Comune di residenza dell'intervistato/a _____ *Codice ISTAT del Comune*_____

*Distretto*_____ *Sezione di censimento*_____

Istruzioni per l'intervistatore:

- le parti scritte in **Arial grassetto** vanno lette all'intervistato
- le parti <u>sottolineate</u> vanno lette con enfasi per attrarre l'attenzione dell'intervistato
- le parti scritte in ***Times New Roman corsivo grassetto*** sono istruzioni speciali per l'intervistatore riferite a quelle domande e non vanno lette all'intervistato

Solo 3 regole:

- **una sola risposta va barrata**, a meno che non sia specificato "***Sono possibili più risposte***"
- si leggono **le domande come** sono **formulate sul questionario**
- **NON** si **leggono le risposte** a meno che non sia specificato "***Leggere le risposte***"

Una facilitazione grafica:

- Ove sono previsti dei **salti che rimandano ad altre pagine del questionario**, perché siano riconosciute più facilmente, le domande richiamate dal salto sono state contrassegnate con una **freccetta (⊢→) posta vicino al numero della domanda**
-  in corrispondenza dei **punti cui fare particolare attenzione** è stato posto un **simbolo**

Questionario PASSI 2011 – Versione 01/01/2011

Data di nascita dell'intervistato/a ____/____/_____ *Sesso dell'intervistato/a* ☐ *M* ☐ *F*

Intervistatore: _____ *Data dell'intervista* ____/____/_____

↳ **Buongiorno, sono** *(nome e cognome di chi parla)*, **La chiamo dalla ASL di _____ per un'intervista sullo stato di salute dei cittadini della quale è stato informato anche il suo medico di famiglia, il Dr.** *Nome* _____ *Cognome* .

↳ **Ha ricevuto per posta la lettera della ASL che Le preannunciava un'intervista telefonica?**

↳ ☐ **SÌ** ☐ **NO**

↳ ***(proseguire comunque)***

Come era scritto nella lettera, la nostra ASL sta facendo delle interviste telefoniche a persone tra i 18 ed i 69 anni.
Le persone sono scelte a caso tra gli assistiti della nostra ASL e lo scopo dell'intervista è conoscere il loro punto di vista su alcuni aspetti che riguardano la salute, per migliorare la qualità dei servizi sanitari offerti.
Lei è una delle persone selezionate per l'intervista che richiede circa 15-20 minuti.
Le informazioni raccolte saranno rese anonime e trattate in base alla legge sulla privacy (D. Lgs. 196/2003). Le ricordo inoltre che può decidere in ogni momento di interrompere l'intervista.
E' disponibile a rispondere ora?
☐ **SÌ** ***(Proseguire con l'intervista)***
☐ **NO**
→ **potrei richiamarla in un momento per lei più opportuno?**
☐ **SÌ (giorno) (ora)**
☐ **NO**
→ **vuole ripensarci dopo aver parlato con il Suo medico?**
☐ **SÌ Bene, allora la richiamo tra qualche giorno. Grazie e a presto.**
☐ **NO**
→ **La ringrazio per l'attenzione che ci ha dedicato. Buongiorno.**

(Se accetta l'intervista)
Mi potrebbe confermare che Lei è nato/a il ____/____/_____ ***(se diverso correggere sopra)***

(*Tagliare e distruggere dopo aver effettuato il caricamento sulla base dati centrale)*

Cognome _____ *Nome* _____

Telefono _____ *Medico di Famiglia*_____

Questionario PASSI 2011 – Versione 01/01/2011

Questionario PASSI 2011 – Versione 01/01/2011

Ora di inizio dell'intervista (ora/min.) | | | | |

SEZIONE 1: Stato di salute e qualità della vita percepita

Le chiederò innanzitutto alcune informazioni generali sul suo stato di salute…

1.1 Come va in generale la sua salute?

Leggere le risposte

☐ Molto bene
☐ Bene
☐ Discretamente
☐ Male
☐ Molto male

Non leggere

☐ Non so

Ora Le farò alcune domande sul suo stato di salute durante gli ultimi 30 giorni.

1.2 Consideri la sua salute fisica, comprese malattie e conseguenze di incidenti. Negli ultimi 30 giorni, per quanti giorni non si è sentito/a bene?

Numero di giorni | | |
☐ Non so / non ricordo

1.3 Adesso pensi agli aspetti psicologici, come problemi emotivi, ansia, depressione, stress. Negli ultimi 30 giorni, per quanti giorni non si è sentito/a bene?

Numero di giorni | | |
☐ Non so / non ricordo

1.4 Ora consideri le sue attività abituali. Negli ultimi 30 giorni, per quanti giorni non è stato/a in grado di svolgerle a causa del cattivo stato di salute fisica o psicologica?

Numero di giorni | | |
☐ Non so / non ricordo

1.5 Un medico le ha mai diagnosticato o confermato una o più delle seguenti malattie?

Leggere le risposte

Insufficienza renale	☐ Sì	☐ No
Bronchite cronica, enfisema, insufficienza respiratoria, asma bronchiale	☐ Sì	☐ No
Ictus o ischemia cerebrale	☐ Sì	☐ No
Infarto del miocardio, ischemia cardiaca o malattia delle coronarie	☐ Sì	☐ No
Altre malattie del cuore ***(es: scompenso valvolopatia)***	☐ Sì	☐ No
Tumori (comprese leucemie e linfomi)	☐ Sì	☐ No
Malattie croniche del fegato, cirrosi	☐ Sì	☐ No

1.6 Negli ultimi 12 mesi, ha fatto la vaccinazione contro l'influenza stagionale?

☐ Sì
☐ No
☐ Non so / non ricordo

(saltare alla Sezione 2: Attività fisica)

1.7 Potrebbe specificarmi in che mese ed anno ha fatto l'ultima vaccinazione contro l'influenza stagionale?

Mese ☐☐ Anno ☐☐☐☐

☐ Non so / non ricordo

SEZIONE 2: Attività fisica

Ora vorrei farle alcune domande sull'attività fisica svolta sia durante sia fuori dal lavoro.

2.1 Lei lavora? *(Si intende lavoro retribuito)*

Leggere le risposte

☐ Sì, in modo continuativo (a tempo pieno o part-time)
☐ Si, ma in modo non continuativo
☐ No

(saltare alla domanda 2.3)

2.2 Durante il suo lavoro, Lei:

Leggere (una sola risposta possibile)

☐ prevalentemente svolge un lavoro pesante che richiede un notevole sforzo fisico (ad. es.: il manovale, il muratore, l'agricoltore)

oppure

☐ prevalentemente cammina o fa lavori che richiedono uno sforzo fisico moderato, (ad. es.: l'operaio in fabbrica, il cameriere, l'addetto alle pulizie)

oppure

☐ prevalentemente sta seduto o in piedi (ad. es.: sta al computer, guida la macchina, fa lavori manuali senza sforzi fisici)

Non leggere

☐ altro

Le faccio adesso qualche domanda sull'attività fisica svolta <u>fuori dal lavoro</u>, sia moderata sia intensa. Cominciamo con quella intensa.

2.3 Negli ultimi 30 giorni, ha svolto qualche attività fisica <u>intensa</u> che provoca grande aumento della respirazione e del battito cardiaco o abbondante sudorazione, come ad esempio correre, pedalare velocemente, fare ginnastica aerobica o sport agonistici?

☐ Sì
☐ No
☐ Non so / non sono sicuro

(saltare alla domanda 2.6)

2.4 Per quanti giorni alla settimana?

Numero di giorni/settimana ☐

☐ Non so / non ricordo

2.5 In questi giorni per quanti minuti in media? ***(riferito ai giorni con attività fisica)***

Numero di minuti/giorno ☐☐☐

☐ Non so / non ricordo

2.6 Negli ultimi 30 giorni, ha svolto qualche attività fisica moderata che comporti un <u>leggero</u> aumento della respirazione e del battito cardiaco o un po' di sudorazione, come ad esempio camminare a passo sostenuto, andare in bicicletta, fare ginnastica dolce, ballare, giardinaggio o lavori in casa come lavare finestre o pavimenti?

☐ Sì

☐ No

☐ Non so / non ricordo } ***(saltare alla domanda 2.9)***

2.7 Per quanti giorni alla settimana?

Numero di giorni ☐☐

☐ Non so / non ricordo

2.8 In questi giorni per quanti minuti in media? ***(riferito ai giorni con attività fisica)***

Numero di minuti/giorno ☐☐☐

☐ Non so / non ricordo

2.9 Secondo lei l'attività fisica che ha fatto negli ultimi 30 giorni è:

Leggere le risposte

☐ Più che sufficiente

☐ Sufficiente

☐ Non proprio sufficiente

☐ Scarsa

Non leggere

☐ Non so

Ora le faccio qualche domanda sui suggerimenti che ha ricevuto <u>negli ultimi 12 mesi</u> sull'attività fisica.

2.10 Negli ultimi 12 mesi Le è stato chiesto da un medico o da un operatore sanitario se fa una regolare attività fisica?

☐ Sì

se la risposta è NO, <u>chiedere</u> se è perché non ha avuto un contatto con un medico o con un operatore sanitario negli ultimi 12 mesi

☐ No, non me lo ha chiesto

☐ No, perché non ho avuto un contatto con medico/operatore sanitario → ***saltare alla Sezione 3: Abitudine al fumo***

☐ Non so / non ricordo

2.11 Negli ultimi 12 mesi Le è stato suggerito da un medico o da un operatore sanitario di fare regolare attività fisica?

☐ Sì

☐ No

☐ Non so / non ricordo

SEZIONE 3: Abitudine al fumo

Ora le farò alcune domande sul fumo di sigaretta.

3.1 Negli ultimi 12 mesi, un medico o un operatore sanitario le ha chiesto se lei è un fumatore?

- ☐ Sì
- ☐ No, non me lo ha chiesto
- ☐ No, perché non ho avuto un contatto con medico/operatore sanitario

(se la risposta è NO, chiedere se è perché non ha avuto un contatto con un medico o con un operatore sanitario negli ultimi 12 mesi)

- ☐ Non so / non ricordo

3.2 In tutta la sua vita, ha fumato in tutto almeno 100 sigarette (5 pacchetti da 20 sigarette)?

- ☐ Sì
- ☐ No
- ☐ Non so / non ricordo

(saltare alla domanda 3.10)

3.3 Attualmente fuma sigarette?

- ☐ Sì
- ☐ No *(saltare alla domanda 3.8)*

3.4 In media quante sigarette fuma al giorno?

Numero ☐☐

- ☐ Meno di una sigaretta al giorno
- ☐ Non so / non ricordo

3.5 Durante gli ultimi 12 mesi, ha smesso di fumare per almeno un giorno nel tentativo di smettere definitivamente?

- ☐ Sì → *(saltare alla domanda 3.7)*
- ☐ No
- ☐ Non so / non ricordo

3.6 Negli ultimi 12 mesi, un medico o un operatore sanitario le ha suggerito di smettere di fumare?

- ☐ Sì, per motivi di salute
- ☐ Sì, a scopo preventivo (in assenza di segni o sintomi)
- ☐ Sì, per tutti e due i motivi

(se la risposta è SI, chiedere per quale motivo: salute, a scopo preventivo, entrambi)

- ☐ No
- ☐ Non so / non ricordo

Saltare alla domanda 3.10

—SOLO PER I FUMATORI CHE HANNO TENTATO DI SMETTERE DI FUMARE —

3.7 L'ultima volta che negli ultimi 12 mesi ha provato a smettere di fumare, come l'ha fatto?

Leggere le risposte (indicare solo il metodo principale)

- ☐ Partecipando ad incontri o corsi che aiutano a smettere di fumare organizzati dalla ASL
- ☐ Partecipando ad incontri o corsi organizzati da altri
- ☐ Prendendo farmaci o cerotti
- ☐ Da solo e per conto mio

(saltare alla domanda 3.7b)

Non leggere

- ☐ Non so / non ricordo
- ☐ Altro

3.7a Quale è stato il motivo principale per cui non ha partecipato a incontri o corsi organizzati dalla ASL?

Leggere le risposte (indicare solo il motivo principale)

- ☐ Non sapevo della loro esistenza
- ☐ Si svolgono in orari o luoghi in cui mi è difficile partecipare
- ☐ E' difficile prenotare/prendere contatto
- ☐ Non credo che servano
- ☐ Costano troppo
- ☐ Non mi risulta che la ASL li organizzi

Non leggere

- ☐ Non so / non ricordo
- ☐ Altro (specificare:……..)

3.7b Negli ultimi 12 mesi, un medico o un operatore sanitario le aveva suggerito di smettere di fumare?

- ☐ Sì, per motivi di salute
- ☐ Sì, a scopo preventivo (in assenza di segni o sintomi)
- ☐ Sì, per tutti e due i motivi

(se la risposta è SI, <u>chiedere</u> per quale motivo: salute, a scopo preventivo, entrambi)

- ☐ No
- ☐ Non so / non ricordo

(saltare alla domanda 3.10)

3.7c Nel suo tentativo di smettere di fumare, quanta influenza ha avuto questo consiglio?

Leggere le risposte

- ☐ Molta
- ☐ Abbastanza
- ☐ Poca
- ☐ Nessuna

(saltare alla domanda 3.10)

——————— *SOLO PER GLI EX-FUMATORI* ———

3.8 Quando ha smesso di fumare?

Leggere le risposte

- ☐ Meno di 6 mesi fa
- ☐ Da 6 mesi a un anno fa
- ☐ Più di un anno fa → ***(saltare alla domanda 3.9c)***

3.8a Negli ultimi 12 mesi, un medico o un operatore sanitario le aveva suggerito di smettere di fumare?

- ☐ Sì, per motivi di salute
- ☐ Sì, a scopo preventivo (in assenza di segni o sintomi)
- ☐ Sì, per tutti e due i motivi

} ***(se la risposta è SI, <u>chiedere</u> per quale motivo: salute, a scopo preventivo, entrambi)***

- ☐ No
- ☐ Non so / non ricordo

} ***saltare alla domanda 3.9***

3.8b Nel riuscire a smettere di fumare, quanta influenza ha avuto questo consiglio?

Leggere le risposte

- ☐ Molta
- ☐ Abbastanza
- ☐ Poca
- ☐ Nessuna

3.9 Come è riuscito/a a smettere di fumare?

Leggere le risposte (indicare solo il metodo principale)

- ☐ Partecipando ad incontri o corsi che aiutano a smettere di fumare organizzati dalla ASL } ***(saltare alla domanda 3.9b)***
- ☐ Partecipando ad incontri o corsi organizzati da altri
- ☐ Prendendo farmaci o cerotti
- ☐ Da solo e per conto mio

Non leggere

- ☐ Non so / non ricordo
- ☐ Altro

3.9a Quale è stato il motivo principale per cui, per smettere di fumare, non ha partecipato a incontri o corsi organizzati dalla ASL?

Leggere le risposte

- ☐ Non sapevo della loro esistenza
- ☐ Si svolgevano in orari o luoghi in cui mi era difficile partecipare
- ☐ Era difficile prenotare/prendere contatto
- ☐ Non credevo che servissero
- ☐ Costavano troppo
- ☐ Non mi risulta che la ASL li organizzasse, quando ho smesso

Non leggere

- ☐ Non so / non ricordo
- ☐ Altro (specificare:……..)

→ **3.9b** In media quante sigarette fumava al giorno?

Numero ☐☐

- ☐ Meno di una sigaretta al giorno
- ☐ Non so / non ricordo

Saltare alla 3.10

----PER COLORO CHE HANNO SMESSO DI FUMARE PIU' DI UN ANNO FA---

→ **3.9c** Come è riuscito/a a smettere di fumare?

Leggere le risposte (indicare solo il metodo principale)

- ☐ Partecipando ad incontri o corsi che aiutano a smettere di fumare organizzati dalla ASL
- ☐ Partecipando ad incontri o corsi organizzati da altri
- ☐ Prendendo farmaci o cerotti
- ☐ Da solo e per conto mio

Non leggere

- ☐ Non so / non ricordo
- ☐ Altro

3.9d In media quante sigarette fumava al giorno?

Numero ☐☐

- ☐ Meno di una sigaretta al giorno
- ☐ Non so / non ricordo

---------PER TUTTI (NON FUMATORI, FUMATORI ED EX-FUMATORI)---------

Vorrei ora chiederle qualcosa sull'esposizione al fumo in casa, nei locali pubblici e sul luogo di lavoro.

3.10 Quale delle seguenti situazioni si avvicina di più alle abitudini sul fumo all'interno di casa sua?

Leggere le risposte

- ☐ Non si fuma in alcuna stanza di casa
- ☐ Si può fumare in alcune stanze o in alcuni orari o situazioni
- ☐ Si può fumare dappertutto

Non leggere

- ☐ Non so / non ricordo

3.11 Nei locali pubblici (come bar, ristoranti, enoteche, pub) che ha frequentato negli ultimi 30 giorni, secondo lei le altre persone:

Leggere le risposte

- ☐ Rispettano sempre i divieti di fumo
- ☐ Li rispettano quasi sempre
- ☐ Li rispettano a volte
- ☐ Non li rispettano mai
- ☐ Non ho frequentato locali pubblici negli ultimi 30 giorni

Non leggere

- ☐ Non so / non ricordo

3.12 Le capita di lavorare in ambienti chiusi? ***(la domanda va somministrata a chi ha risposto che lavora alla domanda 2.1; quindi per chi non lavora barrare "non lavoro")***

- ☐ Sì
- ☐ No
- ☐ Non lavoro

(saltare alla Sezione 4: Alimentazione)

3.13 Nel suo posto di lavoro, le persone con cui lavora e gli eventuali visitatori:

Leggere le risposte

- ☐ Rispettano sempre i divieti di fumo
- ☐ Li rispettano quasi sempre
- ☐ Li rispettano a volte
- ☐ Non li rispettano mai

Non leggere

- ☐ Non so / non ricordo
- ☐ Lavoro da solo

SEZIONE 4: Alimentazione

Passo ora a farle alcune domande sulle sue abitudini alimentari.

4.1 Negli ultimi 12 mesi, un medico o un operatore sanitario le ha suggerito di perdere peso o di mantenere costante il suo peso?

- ☐ Sì
- ☐ No, non me lo ha suggerito
- ☐ No, perché non ho avuto un contatto con medico/operatore sanitario

(se la risposta è NO, chiedere se è perché non ha avuto un contatto con un medico o con un operatore sanitario negli ultimi 12 mesi)

- ☐ Non so / non ricordo

4.2 Attualmente sta seguendo una dieta per perdere o mantenere il suo peso?

- ☐ Sì
- ☐ No

4.3 Secondo lei il suo peso attuale è:

Leggere le risposte

- ☐ Troppo alto
- ☐ Troppo basso
- ☐ Più o meno giusto

Non leggere

- ☐ Non so

4.4 …

Passo ora a chiederle il suo consumo abituale di frutta e verdura. Consideri che per "porzione di frutta o verdura" si intende un quantitativo di frutta o verdura cruda che può essere contenuto sul palmo di una mano, oppure mezzo piatto di verdura cotta.

4.5 … quindi, le chiedo: in una sua giornata tipo, quante porzioni di frutta o verdura mangia?

Leggere le risposte

- ☐ Nessuna
- ☐ 1-2
- ☐ 3-4
- ☐ 5 o più

SEZIONE 5: Assunzione di alcol

Ora vorrei farle qualche domanda sul consumo di alcol.

5.1 Durante gli ultimi 30 giorni, quanti giorni ha bevuto almeno una unità di bevanda alcolica? Per "unità di bevanda alcolica" intendiamo un bicchiere di vino, o una lattina di birra oppure un bicchierino di liquore.

Numero ☐☐
☐ Mai
☐ Non so / non ricordo
(saltare alla domanda 5.10)

5.2 Nei giorni in cui ha bevuto, quante unità di bevande alcoliche ha bevuto in media al giorno?

Numero ☐☐
☐ Non so / non ricordo

5.3 Durante gli ultimi 30 giorni, in quale momento della settimana ha bevuto queste bevande alcoliche?

Leggere le risposte

☐ Prevalentemente nei fine settimana
☐ Prevalentemente nei giorni feriali o durante tutta la settimana

5.4 E quando ha bevuto queste bevande alcoliche rispetto ai pasti?

Leggere le risposte

☐ Solo durante i pasti
☐ Prevalentemente durante i pasti
☐ Prevalentemente fuori dai pasti
☐ Solo fuori dai pasti

5.5

(per gli UOMINI)

Considerando tutti i tipi di bevande alcoliche, negli ultimi 30 giorni quante volte ha bevuto ***5*** o più unità in una unica occasione (ad esempio una serata con amici)?

☐ Numero ☐☐
☐ Mai
☐ Non so / non ricordo

(per le DONNE)

Considerando tutti i tipi di bevande alcoliche, negli ultimi 30 giorni quante volte ha bevuto ***4*** o più unità in una unica occasione (ad esempio una serata con amici)?

☐ Numero ☐☐
☐ Mai
☐ Non so / non ricordo

5.6 Durante gli ultimi 30 giorni le è capitato di guidare un'auto o una moto/scooter dopo aver bevuto, nell'ora precedente, 2 o più unità di una bevanda alcolica?

Leggere le risposte

☐ Si → quante volte? ☐☐
☐ No
☐ Non ho guidato negli ultimi 30 giorni

Non leggere

☐ Non so / non ricordo

5.7 Durante gli ultimi 30 giorni, le è capitato di salire in auto o in moto/scooter con un guidatore che aveva bevuto nell'ora precedente 2 o più unità di una bevanda alcolica?

Leggere le risposte

☐ Si → quante volte? ☐☐

☐ No

☐ Non sono salito su auto o moto negli ultimi 30 giorni

Non leggere

☐ Non so / non ricordo

5.8 Durante gli ultimi 12 mesi, un medico o un operatore sanitario le ha chiesto quanto alcol beve?

☐ Sì

☐ No, non me lo ha chiesto

☐ No, perché non ho avuto un contatto con medico/operatore sanitario

(se la risposta è NO, chiedere se è perché non ha avuto un contatto con un medico o con un operatore sanitario negli ultimi 12 mesi)

☐ Non so / non ricordo

5.9 Durante gli ultimi 12 mesi, qualcuno le ha suggerito di bere meno alcol?

☐ Sì, il medico o un operatore sanitario

☐ Sì, familiari o amici

☐ Sì, altro

(se la risposta è SI, chiedere "Da chi ha avuto il consiglio?" Sono possibili più risposte)

☐ No, non mi è stato consigliato

☐ Non so / non ricordo

(saltare alla domanda 6.1)

— PER CHI DICHIARA DI NON AVER BEVUTO NEGLI ULTIMI 30 GIORNI —

5.10 Durante gli ultimi 30 giorni, le è capitato di salire in auto o in moto/scooter con un guidatore che aveva bevuto nell'ora precedente 2 o più unità di una bevanda alcolica?

Leggere le risposte

☐ Si → quante volte? ☐☐

☐ No

☐ Non sono salito su auto o moto negli ultimi 30 giorni

Non leggere

☐ Non so / non ricordo

5.11 Durante gli ultimi 12 mesi, un medico o un operatore sanitario le ha chiesto quanto alcol beve?

☐ Sì

☐ No, non me lo ha chiesto

☐ No, perché non ho avuto un contatto con medico/operatore sanitario

(se la risposta è NO, chiedere se è perché non ha avuto un contatto con un medico o con un operatore sanitario negli ultimi 12 mesi)

☐ Non so / non ricordo

SEZIONE 6: Sicurezza stradale

Ora le chiederò alcune informazioni sull'uso delle cinture di sicurezza e del casco.

6.1 Quando va in auto, sia in città sia fuori, mette la cintura quando è seduto sui sedili anteriori?

Leggere le risposte

☐ Sempre
☐ Spesso
☐ A volte
☐ Mai ***(Segnare Mai anche se ha l'esenzione)***
☐ Non viaggio mai sui sedili anteriori
☐ Non vado in auto ***(saltare alla domanda 6.3)***

6.2 Quando va in auto, sia in città sia fuori, mette la cintura quando è seduto sui sedili posteriori?

Leggere le risposte

☐ Sempre
☐ Spesso
☐ A volte
☐ Mai ***(Segnare Mai anche se ha l'esenzione)***
☐ Non viaggio mai sui sedili posteriori

6.3 Negli ultimi 12 mesi, è mai salito/a su una motocicletta/scooter/motorino, come guidatore o passeggero?

☐ Sì
☐ No
☐ Non so / non ricordo } ***(saltare alla domanda 6.5)***

6.4 Quando va in motocicletta/scooter/motorino, sia in città sia fuori, mette il casco?

Leggere le risposte

☐ Sempre
☐ Spesso
☐ A volte
☐ Mai

6.5 Negli ultimi 12 mesi, è stato fermato dalle forze dell'ordine (Polizia Stradale, Vigili Urbani, Carabinieri, Polizia di Stato) mentre era <u>alla guida</u> di una macchina o di una moto?

Segnare quanto riportato spontaneamente dall'intervistato

☐ Si → quante volte? ☐☐
☐ No, non sono stato fermato
☐ No, non ho guidato una auto/moto negli ultimi 12 mesi
☐ Non so / non ricordo

} ***saltare alla Sezione 7: Rischio cardiovascolare***

6.6 In occasione di questo/i controllo/i, Le è stato effettuato anche l'etilotest (cioè il cosiddetto "test del palloncino")?

(il test si effettua soffiando in un tubo e serve per valutare se una persona ha bevuto alcol)

☐ Si → quante volte? ☐☐
☐ No
☐ Non so / non ricordo

SEZIONE 7: Rischio cardiovascolare

Vorrei farle ora qualche domanda su esami e farmaci che le sono stati prescritti o consigliati.

7.1 Un medico o altro operatore sanitario le ha mai misurato la pressione arteriosa?

☐ Sì
☐ No
☐ Non so / non ricordo } ***(saltare alla domanda 7.6)***

7.2 Quando è stata l'ultima volta?

Leggere le risposte

☐ Negli ultimi 12 mesi
☐ Tra 1 e 2 anni fa
☐ Più di 2 anni fa

Non leggere

☐ Non so / non ricordo

7.3 Un medico le ha mai detto che Lei è iperteso, cioè che ha la pressione alta?

☐ Sì
☐ No
☐ Non so / non ricordo } ***(saltare alla domanda 7.6)***

7.4 Le è mai stato suggerito da un medico di tenere sotto controllo la sua pressione, attraverso una o più delle seguenti indicazioni?

Leggere tutte le indicazioni

Riduzione del sale nel cibo	☐ Sì	☐ No	☐ Non ricordo
Attività fisica regolare	☐ Sì	☐ No	☐ Non ricordo
Perdita o mantenimento del peso corporeo	☐ Sì	☐ No	☐ Non ricordo

7.5 Prende attualmente farmaci per tenere bassa la pressione?

☐ Sì
☐ No
☐ Non so / non ricordo

Ora vorrei farle qualche domanda sulla misurazione del colesterolo.

7.6 Il colesterolo è un grasso presente nel sangue. Ha mai fatto gli esami per il colesterolo?

☐ Sì
☐ No
☐ Non so / non ricordo } ***(saltare alla domanda 7.10diab1)***

7.7 Quando è stata l'ultima volta?

Leggere le risposte

☐ Negli ultimi 12 mesi
☐ Tra 1 e 2 anni fa
☐ Più di 2 anni fa

Non leggere

☐ Non so / non ricordo

7.8 Un medico le ha mai detto che ha il colesterolo alto?

☐ Sì
☐ No
☐ Non so / non ricordo } ***(saltare alla domanda 7.10diab1)***

7.9 Le è mai stato suggerito da un medico di tenere sotto controllo il livello di colesterolo, attraverso le seguenti indicazioni:

Leggere tutte le indicazioni

Minor consumo di carne e formaggi	☐ Sì	☐ No	☐ Non ricordo
Attività fisica regolare	☐ Sì	☐ No	☐ Non ricordo
Perdita o mantenimento del peso corporeo	☐ Sì	☐ No	☐ Non ricordo
Aumento di frutta e verdura nell'alimentazione	☐ Sì	☐ No	☐ Non ricordo

7.10 Prende attualmente farmaci per tenere basso il colesterolo?

☐ Sì
☐ No
☐ Non so / non ricordo

Ora vorrei farle qualche domanda sul diabete.

7.10diab1 Un medico le ha mai diagnosticato il diabete?

☐ Sì
☐ No } ***(saltare alla domanda 7.10b)***
☐ Non so / non ricordo

7.10diab2 Quando ha saputo per la prima volta di avere il diabete?

Indicare l'anno o l'età a seconda di come ricorda l'intervistato/a

Anno ☐☐☐☐ e/o Età (in anni) ☐☐ Non so/non ricordo ☐

7.10diab3 Da chi è seguito <u>principalmente</u> per il diabete?

Leggere (una sola risposta possibile)

☐ dal medico di famiglia
☐ sia dal medico di famiglia sia dal centro diabetologico
☐ dal centro diabetologico
☐ da un altro medico (specificare ……..….)

Non leggere

☐ da nessuno
☐ non so / non ricordo

} ***(saltare alla domanda 7.10diab6)***

7.10diab4 Negli ultimi 12 mesi quante volte si è rivolto al medico di famiglia per un controllo del diabete?

Numero ☐☐
☐ Mai
☐ Non so / non ricordo

7.10diab5 Negli ultimi 12 mesi quante volte si è rivolto al centro diabetologico per un controllo del diabete?

Numero ☐☐
☐ Mai
☐ Non so / non ricordo

C'è un esame del sangue che si chiama "emoglobina glicosilata", detta anche glicata oppure HbA1c *(leggere: acca-bi-a-uno-ci)*. Questo esame misura il livello medio di glicemia negli ultimi tre mesi.

7.10diab6 Ha mai sentito parlare di questo esame?

☐ Si
☐ No
☐ Non so/non ricordo

} ***(saltare alla domanda 7.10diab8)***

7.10diab7 Ha mai fatto questo esame?

☐ Si → ***Se si, chiedere quando è stata l'ultima volta:***
Leggere le risposte
- ☐ Negli ultimi 4 mesi
- ☐ Tra i 4 e i 12 mesi fa
- ☐ Più di 12 mesi fa
- ☐ Non so/non ricordo

☐ No
☐ Non so/non ricordo

7.10diab8 Prende attualmente farmaci per il diabete?

☐ Si

☐ No
☐ Non so/non ricordo } *(saltare alla domanda 7.10b)*

7.10diab9 Che tipo di farmaci assume per il diabete?

(Non leggere le risposte. Possibile più di una risposta)

☐ Orali (compresse, pillole)
☐ Insulina (iniezioni o microinfusori)
☐ Iniezione di altri farmaci (nome commerciale di exenatide: Byetta)
☐ Non so/non ricordo

7.10b Per le successive domande ho bisogno di chiederle quanti anni ha
(scrivere l'età in anni compiuti) ☐☐

.

se l'intervistato è DONNA con MENO di 25 anni → (saltare alla Sezione 12: Salute mentale)

se l'intervistato è DONNA di 25-34 anni → (saltare alla Sezione 8: Screening Tumore Collo dell'Utero)

se l'intervistato è UOMO con MENO di 35 anni → (saltare alla Sezione 12: Salute mentale)

<u>Pertanto la domanda seguente va somministrata alle persone tra 35 e 69 anni</u>

Alcuni medici hanno cominciato a calcolare, per i loro pazienti, il rischio di avere un infarto o un ictus nei successivi 10 anni. Questo calcolo si chiama "Punteggio" o "Carta del rischio" e si basa sul valore della pressione arteriosa e del colesterolo, sulla presenza di diabete e sull'abitudine al fumo.

7.11 Un medico le ha calcolato questo rischio?

☐ Sì
☐ No
☐ Non so / non ricordo

se l'intervistato è UOMO con MENO di 50 anni → (saltare alla Sezione 12: Salute mentale)

se l'intervistato è UOMO con età compresa tra 50 e 69 anni → (saltare alla Sezione 10: Screening Tumore del Colon-Retto)

SEZIONE 8: Screening Tumore Collo dell'Utero (donne 25-69 anni)

Proseguo ora con qualche domanda sulla prevenzione dei tumori del collo dell'utero per la quale <u>sono disponibili due esami entrambi validi</u>: il <u>Pap-Test</u> oppure il <u>test dell'HPV per la ricerca del papilloma virus</u>. Entrambi questi esami si eseguono attraverso un prelievo fatto con un tampone vaginale.
Ora Le rivolgo alcune domande relative al primo di questi due esami, il <u>Pap-Test</u>.

8.1 Nel corso della sua vita ha <u>eseguito</u> un Pap-Test a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo } *saltare alla domanda 8.9P5*
☐ Ho fatto un tampone, ma non so se Pap-Test o HPV *saltare alla domanda 8.9P11*

8.2 Quando è stata l'ultima volta che ha fatto il Pap-Test a scopo preventivo?

Leggere le risposte

☐ Negli ultimi 12 mesi
☐ Tra 1 e 2 anni fa
☐ Tra 2 e 3 anni fa
☐ Più di 3 anni fa
☐ Non so / non ricordo

Non leggere } ***saltare alla domanda 8.9P5***

8.3 Ha mai ricevuto una lettera dalla ASL che la invitava a fare un Pap-Test?

☐ Sì
☐ No
☐ Non so / non ricordo

8.4 …

8.5 Le è mai stato consigliato da un medico o da un operatore sanitario di fare regolarmente l'esame del Pap-Test a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ No, perché ho già avuto un intervento di isterectomia
(segnare che ha avuto un'isterectomia se l'intervistata lo dichiara spontaneamente)
☐ Non so / non ricordo

8.6 …

8.7 Ha visto o sentito campagne informative o pubblicitarie di promozione del Pap-Test?

☐ Sì
☐ No
☐ Non so / non ricordo

8.8 …

8.9 Ha dovuto pagare per quest'ultimo Pap-test?

Leggere le risposte

☐ Sì, il ticket
☐ Sì, il costo era completamente a mio carico
☐ No, nessuna spesa

Non leggere

☐ Non so / non ricordo

Le rivolgo ora qualche domanda sul secondo tipo di test per la prevenzione dei tumori del collo dell'utero: il test dell'HPV, per la ricerca del papilloma virus, che si esegue sempre con un tampone vaginale.

8.9P1 Nel corso della sua vita ha eseguito un test dell'HPV a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

(se la donna ha:
- MENO di 40 anni, saltare alla Sezione 12: Salute mentale)
- 40 anni o PIÙ saltare alla Sezione 9: Screening Tumore Mammario)

8.9P2 L'ultimo test dell'HPV è stato fatto:

Leggere le risposte

☐ Contemporaneamente a un Pap-Test
☐ Nei due mesi precedenti
☐ Nei due mesi successivi
☐ A distanza di oltre due mesi

Non leggere

☐ Non so / non ricordo

8.9P3 Quando è stata l'ultima volta che ha fatto il test dell'HPV?

☐ Negli ultimi 12 mesi
☐ Tra 1 e 2 anni fa
☐ Tra 2 e 3 anni fa
☐ Più di 3 anni fa

Non leggere

☐ Non so / non ricordo

(se la donna ha:
- MENO di 40 anni, saltare alla Sezione 12: Salute mentale)
- 40 anni o PIÙ saltare alla Sezione 9: Screening Tumore Mammario)

8.9P4 Ha dovuto pagare per questo esame?

Leggere le risposte

☐ Sì, il ticket
☐ Sì, il costo era completamente a mio carico
☐ No, nessuna spesa

Non leggere

☐ Non so / non ricordo

(se la donna ha MENO di 40 anni, saltare alla Sezione 12: Salute mentale)
(se la donna ha 40 anni o PIÙ saltare alla Sezione 9: Screening Tumore Mammario)

-- SOLO PER LE DONNE di 25-69 ANNI CHE NON HANNO FATTO IL PAP TEST MAI O NON LO HANNO FATTO NEGLI ULTIMI 3 ANNI) ---

Le rivolgo ora qualche domanda sul secondo tipo di test per la prevenzione dei tumori del collo dell'utero: il test dell'HPV, per la ricerca del papilloma virus, che si esegue sempre con un tampone vaginale.

8.9P5 Nel corso della sua vita ha eseguito un test dell'HPV a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

- ☐ Sì
- ☐ No — ***saltare alla domanda 8.10***
- ☐ Non so / non ricordo — ***saltare alla domanda 8.10***
- ☐ Ho fatto un tampone vaginale, ma non so se fosse HPV ***saltare alla domanda 8.9P11***

8.9P6 Quando è stata l'ultima volta che ha fatto questo esame?

- ☐ Negli ultimi 12 mesi
- ☐ Tra 1 e 2 anni fa
- ☐ Tra 2 e 3 anni fa
- ☐ Più di 3 anni fa
- ☐ Non so / non ricordo — ***saltare alla domanda 8.10***

8.9P7 Ha mai ricevuto una lettera dalla ASL che la invitava a fare il test dell'HPV?

- ☐ Sì
- ☐ No
- ☐ Non so / non ricordo

8.9P8 Le è mai stato consigliato da un medico o da un operatore sanitario di fare il test dell'HPV a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

- ☐ Sì
- ☐ No
- ☐ No, perché ho già avuto un intervento di isterectomia

(segnare che ha avuto un'isterectomia se l'intervistata lo dichiara spontaneamente)

- ☐ Non so / non ricordo

8.9P9 Ha visto o sentito campagne informative o pubblicitarie di promozione del test dell'HPV?

- ☐ Sì
- ☐ No
- ☐ Non so / non ricordo

8.9P10 Ha dovuto pagare per questo esame?

Leggere le risposte

- ☐ Sì, il ticket
- ☐ Sì, il costo era completamente a mio carico
- ☐ No, nessuna spesa

Non leggere

- ☐ Non so / non ricordo

(se la donna ha MENO di 40 anni, saltare alla Sezione 12: Salute mentale)
(se la donna ha 40 anni o PIÙ saltare alla Sezione 9: Screening Tumore Mammario)

-SOLO PER LE DONNE DI 25-69 ANNI CHE HANNO EFFETTUATO UN TAMPONE VAGINALE A SCOPO PREVENTIVO, MA NON SANNO SE PAP-TEST O HPV–

8.9P11 Quando è stata l'ultima volta che ha fatto un tampone vaginale a scopo preventivo?

☐ Negli ultimi 12 mesi
☐ Tra 1 e 2 anni fa
☐ Tra 2 e 3 anni fa
☐ Più di 3 anni fa
☐ Non so / non ricordo } ***saltare alla domanda 8.10***

8.9P12 Ha mai ricevuto una lettera dalla ASL che la invitava a fare un Pap-Test o un test dell'HPV?

☐ Sì
☐ No
☐ Non so / non ricordo

8.9P13 Le è mai stato consigliato da un medico o da un operatore sanitario di fare regolarmente questi esami a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ No, perché ho già avuto un intervento di isterectomia
(segnare che ha avuto un'isterectomia se l'intervistata lo dichiara spontaneamente)
☐ Non so / non ricordo

8.9P14 Ha visto o sentito campagne informative o pubblicitarie di promozione per questi esami?

☐ Sì
☐ No
☐ Non so / non ricordo

8.9P15 Ha dovuto pagare per questo ultimo test eseguito a scopo preventivo?

Leggere le risposte

☐ Sì, il ticket
☐ Sì, il costo era completamente a mio carico
☐ No, nessuna spesa

(se la donna ha MENO di 40 anni, saltare alla Sezione 12: Salute mentale)
(se la donna ha 40 anni o PIÙ saltare alla Sezione 9: Screening Tumore Mammario)

-SOLO PER LE DONNE DI 25-69 ANNI CHE NON HANNO MAI EFFETTUATO UN ESAME PREVENTIVO (né PAP-TEST né HPV) O CHE L'HANNO EFFETTUATO PIU' DI TRE ANNI FA -

8.10 Ha mai ricevuto una lettera dalla ASL che la invitava a fare un Pap-Test o un test dell'HPV?

☐ Sì
☐ No
☐ Non so / non ricordo

8.11 Le è mai stato consigliato da un medico o da un operatore sanitario di fare regolarmente questi esami a scopo preventivo, cioè in assenza di sintomi?

☐ Sì
☐ No
☐ No, perché ho già avuto un intervento di isterectomia
(segnare che ha avuto un'isterectomia se l'intervistata lo dichiara spontaneamente)
☐ Non so / non ricordo

8.12 Ha visto o sentito campagne informative o pubblicitarie di promozione per questi esami?

☐ Sì
☐ No
☐ Non so / non ricordo

Se la donna non ha MAI fatto un Pap-Test o un test HPV, non leggere la parola "recentemente"

8.13 Quale è stato il motivo principale per cui **non** ha fatto (recentemente) un Pap test o un test HPV?

Non leggere le risposte (è possibile una sola risposta)

☐ Penso di non averne bisogno
☐ Mi sento imbarazzata/mi vergogno
☐ E' fastidioso/doloroso
☐ Ho paura dei risultati dell'esame
☐ Per pigrizia
☐ Sono già stata operata / per altri motivi sanitari
☐ Non so / non ricordo
☐ Nessuno me l'ha consigliato
☐ Ho trovato difficile contattare l'ASL per informazioni/appuntamento
☐ La sede/data/orario assegnati per l'esame non mi andavano bene
☐ Non ho ricevuto una convocazione
☐ Non ho avuto tempo
☐ Non sono in età target
☐ Altro (specificare …….....)

(se la donna ha MENO di 40 anni, saltare alla Sezione 12: Salute mentale)

SEZIONE 9: Screening Tumore Mammario (donne 40-69 anni)

Vorrei farle una serie di domande sulla mammografia, che è una radiografia fatta al seno per ricercare l'eventuale presenza di tumore alla mammella.

9.1 Nel corso della sua vita ha mai fatto una mammografia a scopo preventivo cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

Se la donna ha 40-49 anni, saltare alla Sezione 12: Salute mentale
Se la donna ha 50-69 anni, saltare alla domanda 9.11

9.2 A che età ha fatto la sua prima mammografia a scopo preventivo?

☐☐ anni
☐ Non so / non ricordo

9.3 Quando è stata l'ultima volta che ha fatto una mammografia a scopo preventivo?

Leggere le risposte

☐ Negli ultimi 12 mesi
☐ Tra 1 e 2 anni fa
☐ Più di 2 anni fa
Non leggere
☐ Non so / non ricordo

Se la donna ha 40-49 anni, saltare alla Sezione 12: Salute mentale
Se la donna ha 50-69 anni, saltare alla domanda 9.11

9.4 Ha mai ricevuto una lettera dalla ASL per fare una mammografia?

☐ Sì
☐ No
☐ Non so / non ricordo

9.5 ….

9.6 Le è mai stato consigliato da un medico o da un operatore sanitario di fare regolari mammografie a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

9.7 ….

9.8 Ha visto o sentito campagne informative o pubblicitarie di promozione della mammografia?

☐ Sì
☐ No
☐ Non so / non ricordo

9.9 ….

9.10 Ha dovuto pagare per quest'ultimo esame?

Leggere le risposte

☐ Sì, il ticket
☐ Sì, il costo era completamente a mio carico
☐ No, nessuna spesa
Non leggere
☐ Non so / non ricordo

(se la donna ha 40-49 anni, saltare alla Sezione 12: Salute mentale)
(se la donna ha 50-69 anni, saltare alla Sezione 10: Screening Tumore del Colon-Retto)

------SOLO PER LE DONNE DI 50-69 ANNI CHE NON HANNO MAI EFFETTUATO UNA MAMMOGRAFIA PREVENTIVA O CHE L'HANNO EFFETTUATA PIU' DI DUE ANNI FA----

9.11 Ha mai ricevuto una lettera dalla ASL per fare una mammografia?

☐ Sì
☐ No
☐ Non so / non ricordo

9.12 Le è mai stato consigliato da un medico o da un operatore sanitario di fare regolari mammografie a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

9.13 Ha visto o sentito campagne informative o pubblicitarie di promozione della mammografia?

☐ Sì
☐ No
☐ Non so / non ricordo

Se la donna non ha MAI fatto una mammografia, non leggere la parola "recentemente"

9.14 Quale è stato il motivo principale per cui **non** ha fatto (recentemente) la mammografia?

Non leggere le risposte (è possibile una sola risposta)

☐ Penso di non averne bisogno
☐ Mi sento imbarazzata/mi vergogno
☐ E' fastidioso/doloroso
☐ Ho paura dei risultati dell'esame
☐ Per pigrizia
☐ Non ho avuto tempo
☐ Nessuno me l'ha consigliato
☐ Ho trovato difficile contattare l'ASL per informazioni/appuntamento
☐ La sede/data/orario assegnati per l'esame non mi andava bene
☐ Non ho ricevuto una convocazione
☐ Altro (specificare .....)
☐ Sono già stata operata / per altri motivi sanitari
☐ Non so / non ricordo

SEZIONE 10: Screening Tumore del Colon-Retto (persone 50-69 anni)

Vorrei ora farle qualche domanda sugli esami per la prevenzione del tumore del colon-retto. Esiste un esame di laboratorio, chiamato "ricerca del sangue occulto", per controllare la presenza di sangue non visibile nelle feci, che prevede la raccolta di feci in un contenitore, anche presso la propria abitazione.

10.1. Ha mai fatto la ricerca del sangue occulto nelle feci a scopo preventivo, cioè in assenza di sintomi o disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

} ***(saltare alla domanda 10.10)***

10.2. Quando è stata l'ultima volta che ha fatto questo esame a scopo preventivo?

Leggere le risposte

☐ Negli ultimi 12 mesi
☐ Tra 1 e 2 anni fa
☐ Più di 2 anni fa

Non leggere

☐ Non so / non ricordo

} ***(saltare alla domanda 10.10)***

10.3. Ha mai ricevuto una lettera dalla sua ASL per fare la ricerca del sangue occulto nelle feci?

☐ Sì
☐ No
☐ Non so / non ricordo

} ***saltare alla domanda 10.5***

10.4. …

10.5. Le è mai stato consigliato da un medico o da un operatore sanitario di fare questo esame a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

} ***saltare alla domanda 10.7***

10.6. …

10.7. Ha visto o sentito campagne informative o pubblicitarie di promozione della ricerca di sangue occulto nelle feci?

☐ Sì
☐ No
☐ Non so / non ricordo

} ***saltare alla domanda 10.9***

10.8. …

10.9. Ha dovuto pagare per quest'ultimo esame?

Leggere le risposte

☐ Sì, il ticket
☐ Sì, il costo era completamente a mio carico
☐ No, nessuna spesa

Non leggere

☐ Non so / non ricordo

(Saltare alla domanda 10.14)

---SOLO PER I 50-69ENNI CHE NON HANNO MAI EFFETTUATO LA RICERCA DEL SANGUE OCCULTO NELLE FECI PREVENTIVA O CHE L'HANNO EFFETTUATA PIU' DI DUE ANNI FA---

10.10. Ha mai ricevuto una lettera dalla sua ASL per fare la ricerca del sangue occulto nelle feci?

☐ Sì
☐ No
☐ Non so / non ricordo

10.11. Le è mai stato consigliato da un medico o da un operatore sanitario di fare questo esame a scopo preventivo, cioè in assenza di sintomi o altri disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

10.12. Ha visto o sentito campagne informative o pubblicitarie di promozione della ricerca di sangue occulto nelle feci?

☐ Sì
☐ No
☐ Non so / non ricordo

Se la persona non ha MAI fatto la ricerca del sangue occulto, non leggere la parola"recentemente"

10.13. Quale è stato il motivo principale per cui **non** ha fatto (recentemente) la ricerca del sangue occulto nelle feci a scopo preventivo

Non leggere le risposte (è possibile una sola risposta)

☐ Penso di non averne bisogno
☐ Mi sento imbarazzato/a / mi vergogno
☐ E' fastidioso
☐ Ho paura dei risultati dell'esame
☐ Per pigrizia
☐ Non ho avuto tempo
☐ Nessuno me l'ha consigliato
☐ Ho trovato difficile contattare l'ASL per informazioni/appuntamento
☐ La sede/data/orario per prendere/ consegnare il campione non mi andava bene
☐ Non ho ricevuto una convocazione
☐ Altro (specificare .....)
☐ Sono già stata operato/a / per altri motivi sanitari
☐ Ho già fatto / mi hanno consigliato di fare la colonscopia/rettosigmoidoscopia
☐ Non so / non ricordo

----------PER TUTTE LE PERSONE DI 50-69 ANNI----------

Ci sono altri esami, fatti con una sonda inserita nell'intestino, che si eseguono per la prevenzione di questo tipo di tumore. Questi esami si chiamano colonscopìa o anche retto-sigmoidoscopìa.

10.14. Ha mai <u>fatto</u> la colonscopìa o la retto-sigmoidoscopìa a scopo preventivo, cioè in assenza di sintomi o disturbi?

☐ Sì
☐ No
☐ Non so / non ricordo

} ***Saltare alla Sezione 12: Salute mentale***

10.15. Quando è stata l'ultima volta che ha fatto questo esame a scopo preventivo?

Leggere le risposte

☐ Negli ultimi 12 mesi
☐ Tra 1 e 5 anni fa
☐ Tra 5 e 10 anni fa
☐ Più di 10 anni fa

Non leggere

☐ Non so / non ricordo

10.16. Ha dovuto pagare per quest'ultimo esame?

Leggere le risposte

☐ Sì, il ticket
☐ Sì, il costo era completamente a mio carico
☐ No, nessuna spesa

Non leggere

☐ Non so / non ricordo

SEZIONE 12: Salute mentale

Vorrei ritornare su un argomento già trattato e farle qualche domanda su <u>come lei si sente dal punto di vista psicologico</u> e su come affronta la vita di ogni giorno.

12.1 Nelle ultime <u>2 settimane</u>, per quanti giorni ha provato poco interesse o piacere nel fare le cose?

Numero di giorni ☐☐ (0-14)
☐ Non so / non ricordo

12.2 Nelle ultime <u>2 settimane</u>, per quanti giorni si è sentito/a giù di morale, depresso/a o senza speranze?

Numero di giorni ☐☐ (0-14)
☐ Non so / non ricordo

Per quelli che rispondono <u>7 giorni o più</u>, <u>a una delle domande o a entrambe</u>, proseguire con la domanda 12.3, altrimenti saltare alla Sezione 13 – Sicurezza Domestica

12.3 A causa di questi problemi, anche in passato, si è rivolto ad una o più delle seguenti persone? ***(Sono possibili più risposte)***

Leggere le risposte

☐ Personale sanitario (medico di famiglia, psicologo, infermiere …)
☐ Persone di fiducia (familiari, amici)
☐ No, a nessuno

Non leggere

☐ Non so / non ricordo

SEZIONE 13: Sicurezza domestica

Vorrei ora farle qualche domanda sugli infortuni domestici, cioè quelli che avvengono sia in casa, sia negli ambienti esterni quali giardino, garage, cantina, terrazzo.

13.1 Secondo lei, quale è la possibilità <u>per il suo nucleo familiare</u> di avere un infortunio in ambiente domestico?

Leggere le risposte

☐ Assente
☐ Bassa
☐ Alta
☐ Molto alta

13.1b Negli ultimi 12 mesi, Lei ha avuto un infortunio domestico per il quale è dovuto ricorrere alle cure del medico di famiglia, del Pronto Soccorso o dell'Ospedale?

☐ Sì
☐ No
☐ Non so / non ricordo

SEZIONE 14: Dati socio – anagrafici

14.1 Quale è il suo attuale stato civile?

☐ Coniugato
☐ Celibe/nubile
☐ Vedovo/a
☐ Separato/a-divorziato/a

(anche se convivente)

14.2 Chi abita in casa con Lei? ***(sono possibili più risposte)***

Leggere le risposte

☐ Nessuno, vivo da solo ⟶ ***(saltare alla domanda 14.3)***

☐ Coniuge/Compagno/a

☐ Figli/Bambini fino a 14 anni ***(<u>Attenzione</u>: per chi risponde Sì, ricordarsi delle domande alla fine del questionario)***

☐ Figli, altri parenti o amici tra 15–64 anni
☐ Altri parenti o amici di 65 anni e più

14.3 Qual è la sua cittadinanza?

☐ Italiana ***(saltare alla domanda 14.4)***
☐ Straniera
☐ Doppia (italiana e straniera)

(<u>specificare: …….....…..</u>)

14.3a Da quanto tempo vive in Italia?

Numero anni ☐☐
☐ Meno di un anno
☐ Non so / non ricordo

14.4 Quale è il suo titolo di studio?

☐ Nessun titolo
☐ Licenza elementare
☐ Licenza di scuola media
☐ Diploma o qualifica di scuola media superiore
☐ Laurea/Diploma universitario o titolo superiore

14.5 Con le risorse finanziarie a sua disposizione (da reddito proprio o familiare) come arriva a fine mese?

Leggere le risposte

☐ Molto facilmente
☐ Abbastanza facilmente
☐ Con qualche difficoltà
☐ Con molte difficoltà

14.6 Può dirmi la sua altezza (senza scarpe)?

☐☐☐ cm

14.7 Può dirmi il suo peso, senza scarpe ed abiti o con abiti leggeri?

☐☐☐ Kg

14.8 Rispetto ad un anno fa, il suo peso è cambiato oppure è stabile?

Leggere le risposte

☐ aumentato (almeno 2 chili in più)
☐ stabile
☐ diminuito (almeno 2 chili in meno)

Non leggere

☐ sono/ero in gravidanza
☐ Non so / non ricordo

Se alla domanda 14.2 l'intervistato ha risposto che <u>abita con persone fino a 14 anni (compresi)</u> passare alla domanda 14.9, altrimenti saltare alla frase di congedo.
Visto che vive con bambini/ragazzi, le chiedo:

14.9 Quanti anni ha il/la bambino/a-ragazzo/a più piccolo/a (se sono più di uno)?

☐ Età in anni: ☐☐ ***(se il bambino ha 7 anni o più, saltare alla frase di congedo)***
☐ Non so / non ricordo

14.10 Quando il/la piccolo/a viaggia con lei in auto ha difficoltà a farlo/a stare seduto/a ed allacciato/a al seggiolino?

Leggere le risposte

☐ No, nessuna difficoltà
☐ Sì, qualche difficoltà
☐ Sì, molte difficoltà
☐ Non ho il seggiolino, non lo uso
☐ Non vado mai in auto con il/la bambino/a

Non leggere

☐ Non so / non ricordo

14.11 …

14.12 Negli ultimi 12 mesi, ha visto o sentito campagne informative o pubblicitarie di promozione dei dispositivi di sicurezza per i bambini in automobile?

☐ Sì
☐ No
☐ Non so / non ricordo

Abbiamo finito. La ringrazio moltissimo per la collaborazione e la disponibilità…

Fine intervista (ora/min.) | | | | |

BIBLIOGRAFIA

[Azzalini 2001] A. Azzalini (2001), *Inferenza statistica - Una presentazione basata sul concetto di verosimiglianza*, 2^a edizione, Springer - Verlag Italia, Milano.

[Di Fonzo, Lisi 2005] T. Di Fonzo, F. Lisi (2005), *Serie storiche economiche - Analisi statistiche e applicazioni*, Carocci editore, Roma.

[Gelman, Hill 2006] A. Gelman, J. Hill (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press, Cambridge.

[Goldstein 1999] H. Goldstein (1999), *Multilevel statistical models*, London: Institute of education, Multilevel models project.

[Passi] `http://www.epicentro.iss.it/passi/`.

[Pezzato 2010] L. Pezzato (2010), Modelli di regressione con dati longitudinali, *Tesi di laurea triennale in Statistica e informatica per la gestione delle imprese*, Università Ca' Foscari, Venezia.

[Ricci 2005] V. Ricci (2005), *Analisi delle serie storiche con R*, `http://cran.r-project.org/doc/contrib/Ricci-ts-italian.pdf`.

[Ventura 2008] L. Ventura (2008), Appunti, *Modelli statistici II - Modelli lineari generalizzati*, `http://www.statistica.unipd.it/insegnamenti/modstat2/matdid/lucidilez4_09.pdf`.