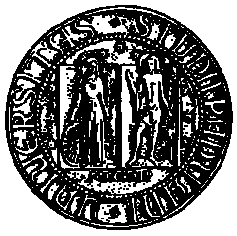


UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

TESI DI LAUREA

LO STUDIO DEL CASO "BUTTERFLY BALLOT" DELLE ELEZIONI AMERICANE DEL 2000 NELLA CONTEA DI PALM BEACH

RELATORE: CH.MO PROF. STUART COLES

LAUREANDO: FILIPPO DA RE

ANNO ACCADEMICO 2005/2006

Dedicato alla
squadra Palm Beach.
Un grazie a Bissa e Mauro
per la scelta del titolo
e a Marta
per la pazienza
e l'aiuto.

INDICE

1. INTRODUZIONE

Il lavoro che segue presenta un’analisi statistica su di un avvenimento registrato durante le elezioni presidenziali negli Stati Uniti d’America.

Come noto, negli USA, non sono i cittadini ad eleggere direttamente il presidente ma 538 cosiddetti "grandi elettori". I cittadini votanti, sulla scheda esprimono la preferenza per un candidato presidente, ma in realtà eleggono una lista di "grandi elettori" associati con lui.
I voti dei cittadini (detti "voti popolari") si contano Stato per Stato e non a livello nazionale. Colui che vince - anche di un solo voto - in uno Stato, si prende tutti i "grandi elettori" in palio in quello Stato e chi riesce a far eleggere almeno 270 grandi elettori (il 50%+1) diviene presidente. Il numero di elettori è determinato nella seguente maniera: ogni Stato, piccolo o grande, ha diritto a due grandi elettori più tanti altri quanti sono i deputati inviati alla Camera dei rappresentanti; i deputati alla Camera sono attribuiti grossomodo in proporzione alla popolazione.

Durante le elezioni presidenziali del 2000, i maggiori candidati in corsa per la Casa Bianca erano il democratico Al Gore e il repubblicano George W. Bush. I due candidati si spartirono l’america in parti eguali fino ad arrivare all’ultimo stato mancante, la Florida, che, con i suoi 25 seggi, avrebbe designato il vincitore della campagna politica. Ufficialmente, dopo l’intervento della Corte Suprema, Bush vinse le elezioni in Florida con uno scarto di 537 voti, su di un totale di 6 milioni, sull’avversario Gore. Ma si accesero subito delle polemiche. Molti, infatti, sostennero che le macchine punzonatrici non funzionarono a dovere, altri, invece, sostennero che le schede a farfalla (da qui il termine che contraddistinse il caso: “butterfly ballot”) misero confusione ai votanti. In particolare, lasciò perplesso quanto accadde nella contea di Palm Beach, la terza (dopo Miami-Dade e Broward) per importanza con più di un milione d’abitanti e con 430 mila voti registrati. Questa contea, da sempre democratica e con Gore che registrò più del 62% di voti a suo favore, riportò 3407 schede a favore di Pat Bachanan, il candidato del partito riformista. Questo dato suggerì il possibile confondimento nel voto visto che Buchanan in tutto lo stato non era mai andato oltre i mille voti.

Sono stati fatti molti studi sull'accaduto, alcuni per dimostrare l'erroneità dei risultati ottenuti, altri per confermare la vittoria di Bush, altri ancora per sostenere l'ipotetica vittoria di Gore a cui,secondo alcuni ricercatori, apparterrebbero i voti "in eccesso" di Buchanan; con questo lavoro, invece, si cercherà semplicemente di valutare statisticamente il caso e, tramite la stima di un modello lineare generalizzato, provare a prevedere il risultato che avrebbe dovuto ottenere Pat Buchanan nella contea di Palm Beach, per cercare di capire se i 3407 voti conteggiati a favore del candidato riformista,sono, oppure no, un outlier, un valore erratico.

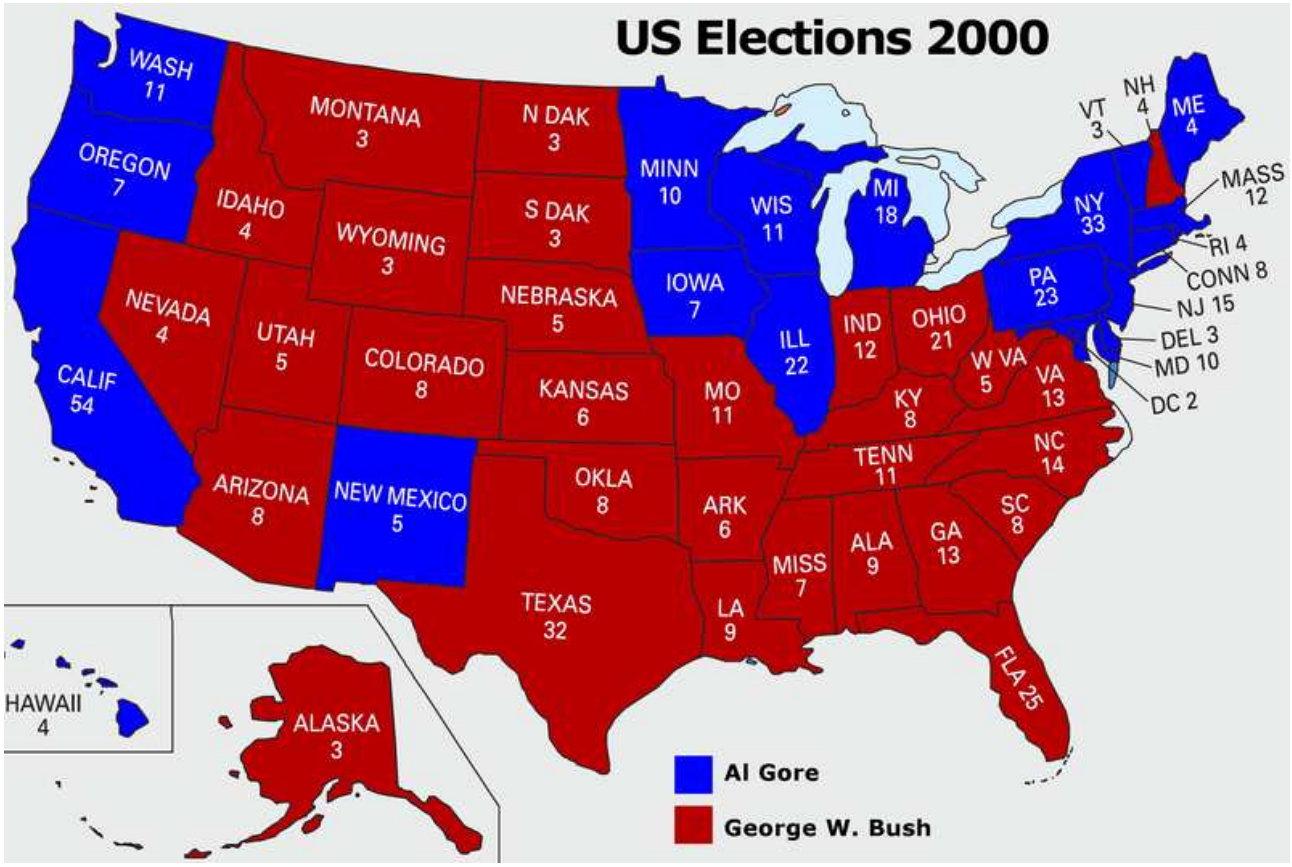


Fig.1: Risultati delle elezioni del 2000 con 271 grandi elettori per Bush, contro i 267 di Gore, a dimostrare l'importanza decisiva della Florida per la vittoria finale del candidato repubblicano

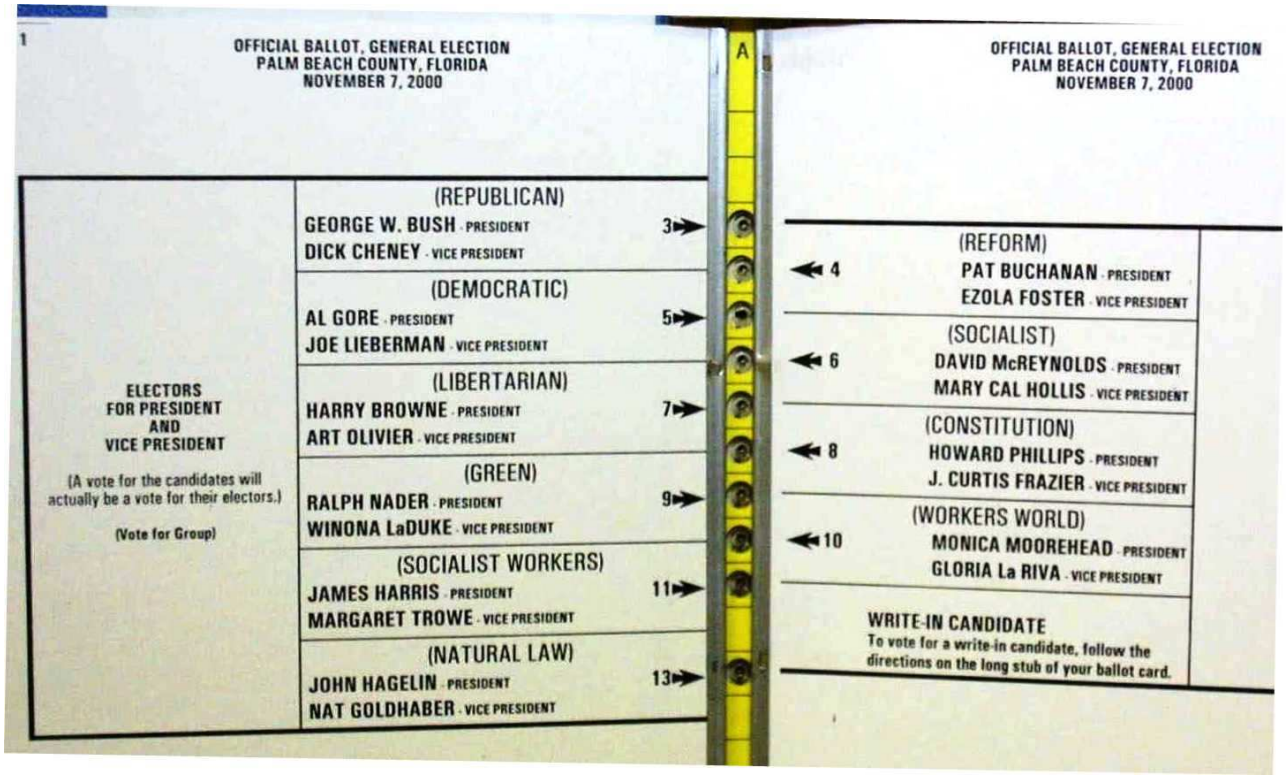


Fig.2: Immagine di una scheda elettorale della contea di Palm Beach

2. DATI UTILIZZATI

2.1 Descrizione delle varibili

In questo studio sono stati impiegati due data sets differenti assemblati tra loro per comodità. Il primo si riferisce a dati demografici compilati dall'ufficio dell'U.S. Census, il secondo è l'elenco dei risultati elettorali ottenuti nelle varie contee della Florida (forniti dal Florida Division of Elections). Tranne che per le variabili "co" (codice identificativo dall'1 al 67, uno per contea in ordine alfabetico) e "lon" e "lat" (coordinate di longitudine e latitudine del centro d'ogni contea), le restanti variabili demografiche, poi riportate in tabella, sono:

- Pop: numero della popolazione nella contea, 1997 ("npop")
- Whi: percentuale di bianchi " " , 1996 ("whit")
- Bla: percentuale di neri " " , 1996 ("blac")
- Hisp: percentuale di ispanici " " , 1996 ("hisp")

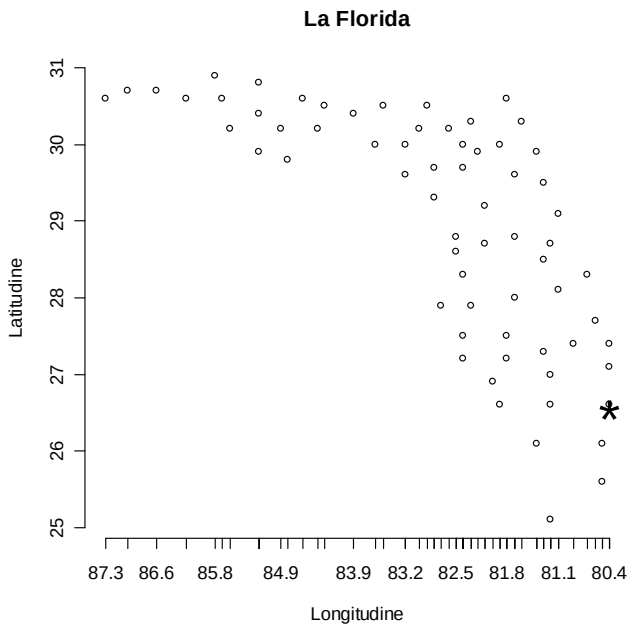
(nota: il totale tra bianchi, neri ed ispanici può superare il 100% visto che nell'ultima categoria possono essere state catalogate altre razze)

- ≥65: percentuale della popolazione di età maggiore o uguale ai 65 anni, 1996 (“o65”)
- HS: percentuale della popolazione che ha terminato l’high school, 1990 (“hsed”)
- Coll: percentuale della popolazione che ha terminato il college, 1990 (“coll”)
- Inc: reddito personale medio, 1994 (“inco”)

Le variabili riguardanti i risultati elettorali, in numero totale di voti ottenuti, sono:

- Bush: George W. Bush del partito Repubblicano (“bush”)
- Gore: Al Gore del partito Democratico (“gore”)
- Brow: Harry Browne del Libertarian Party (“brow”)
- Nade: Ralph Nader del Green Party (“nade”)
- Har: James Harris del Socialist Workers Party (“harr”)
- Hag: John Hagelin del Natural Law Party (“hage”)
- Buc: Pat Buchanan del Reform Party (“buch”)
- Mc: David McReynolds del Socialist Party (“mcre”)
- Ph: Howard Phillips del Constitution Party (“phil”)
- Mo: Monica Moorehead del Workers World Party (“moor”)

Precisiamo che, tra tutti questi candidati, Bush e Gore hanno ottenuto quasi il 98% di tutti i voti, i rimanenti candidati hanno preso il restante 2%.



Oltre all’anno di raccolta dei dati, per le variabili demografiche, è stato messo tra parentesi il nome con il quale sono state etichettate per l’analisi in R.
Nella cartina è segnata la contea di Palm Beach.
Nelle tabelle che seguono sono state elencate tutte le variabili sopra descritte per tutte le 67 contee della Florida.

County	Pop	Whi	Bla	Hisp	≥ 65	HS	Coll	Inc
Alachua	198326	74.4	21.8	4.7	9.4	82.7	34.6	19412
Baker	20761	82.4	16.8	1.5	7.7	64.1	5.7	14859
Bay	146223	84.2	12.4	2.4	11.9	74.7	15.7	17838
Bradford	24646	76.1	22.9	2.6	11.8	65.0	8.1	13681
Brevard	460977	88.3	9.2	4.1	16.5	82.3	20.4	19567
Broward	1470758	80.3	17.5	10.9	20.3	76.8	18.8	24706
Calhoun	12337	81.6	16.9	1.6	14.3	55.9	8.2	12570
Charlotte	133681	94.3	4.4	3.4	33.4	75.7	13.4	18977
Citrus	112454	96.2	2.8	2.5	30.7	68.6	10.4	16060
Clay	135179	91.0	6.0	3.5	7.9	81.2	17.9	18598
Collier	195731	93.3	5.7	17.1	21.5	79.0	22.3	30906
Columbia	52856	78.3	20.5	1.9	12.3	69.0	11.0	15349
Desoto	26259	80.6	18.1	12.1	18.0	54.5	7.6	16544
Dixie	12563	89.8	9.5	1.2	14.4	57.7	6.2	12035
Duval	732622	69.4	27.5	3.4	10.7	76.9	18.4	20686
Escambia	282604	73.3	22.7	2.6	11.7	76.2	18.2	17661
Flagler	46128	88.5	9.8	5.9	23.0	78.7	17.3	15613
Franklin	10133	84.5	14.5	1.0	17.8	59.5	12.4	15735
Gadsden	45441	37.6	61.8	2.9	11.6	59.9	11.2	14416
Gilchrist	13367	90.0	9.3	2.1	13.0	63.0	7.4	12865
Glades	9698	79.6	13.7	10.1	15.3	57.4	7.1	14789
Gulf	13926	73.9	25.2	1.1	13.6	66.4	9.2	15482
Hamilton	12521	56.3	43.0	3.6	10.9	58.4	7.0	12357
Hardee	22113	93.1	5.9	28.4	13.3	54.8	8.6	16812
Hendry	31634	78.2	18.8	26.6	9.9	56.6	10.0	17823
Hernando	125537	94.4	4.6	4.0	29.6	70.5	9.7	16062
Highlands	76854	87.1	11.6	6.7	32.4	68.2	10.9	17655
Hillsborough	909444	82.8	14.9	16.0	12.3	75.6	20.2	20167
Holmes	18382	91.7	6.5	1.7	15.5	57.1	7.4	12790
Indian River	99215	89.2	9.9	3.9	26.6	76.5	19.1	28977
Jackson	45706	69.5	29.6	3.5	14.4	61.6	10.9	15519
Jefferson	13232	49.4	50.1	1.3	13.4	64.1	14.7	15574
Lafayette	6289	83.0	16.4	5.1	10.7	58.2	5.2	13663
Lake	196214	88.2	10.9	3.8	26.3	70.6	12.7	18269
Lee	387091	91.1	7.8	5.9	24.4	76.9	16.4	22053
Leon	215170	70.4	27.3	3.1	8.4	84.9	37.1	16705
Levy	32254	84.4	14.2	2.6	17.6	62.8	8.3	13745
Liberty	6703	78.1	20.9	3.1	10.7	56.7	7.3	14896

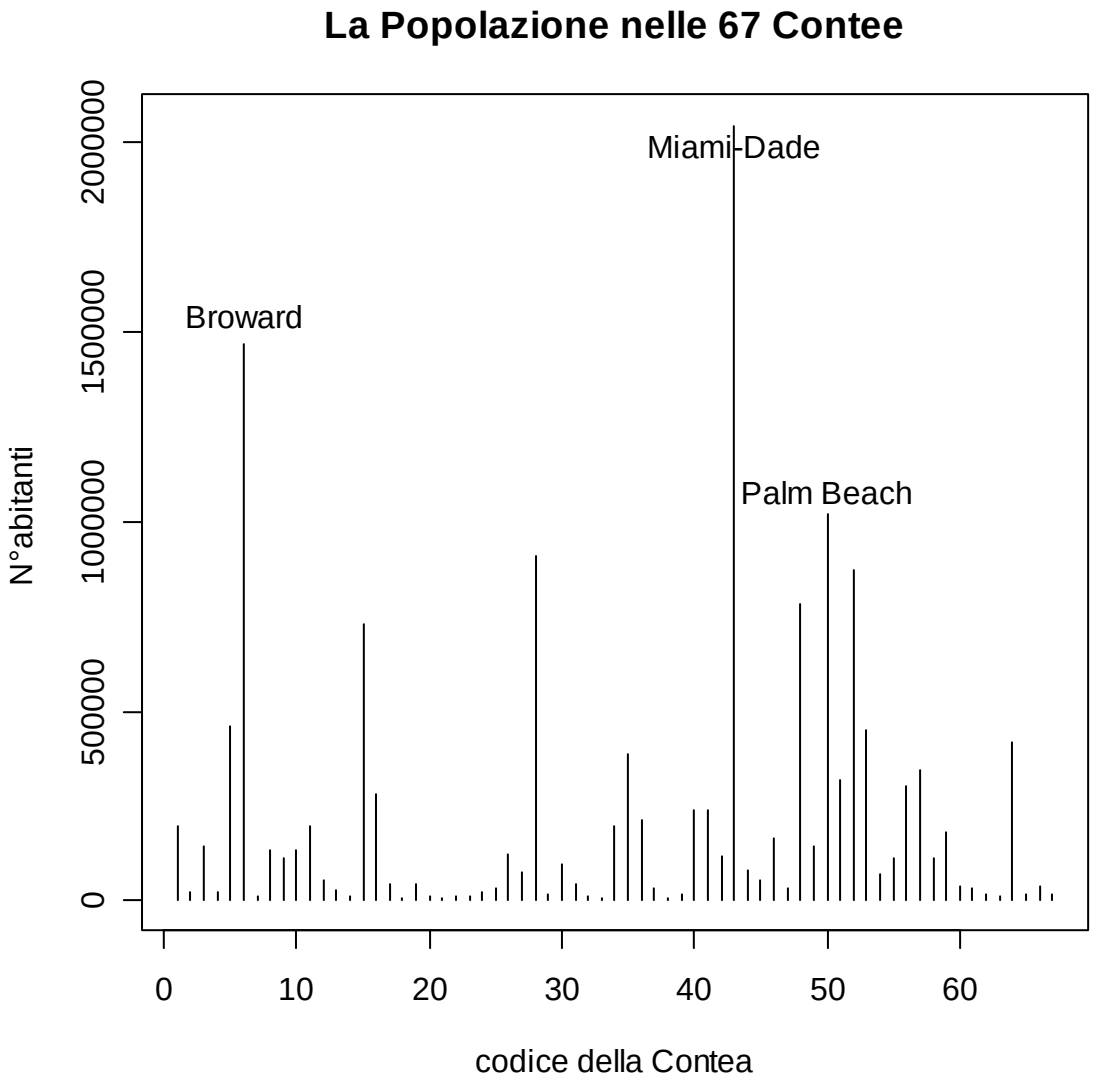
County	Pop	Whi	Bla	Hisp	≥ 65	HS	Coll	Inc
Madison	17558	53.9	45.6	1.9	13.8	56.5	9.7	13002
Manatee	237159	89.8	9.0	5.8	27.8	75.6	15.5	23031
Marion	237308	84.3	14.6	4.0	21.4	69.6	11.5	14502
Martin	116087	91.8	6.9	6.2	26.6	79.7	20.3	31996
Miami-Dade	2044600	77.0	21.2	54.4	14.4	65.0	18.8	20014
Monroe	81919	92.3	6.2	15.8	15.9	79.7	20.3	25160
Nassau	54096	87.3	11.9	1.5	9.8	71.2	21.5	20874
Okaloosa	167580	85.3	10.3	4.2	9.1	83.8	21.0	18959
Okeechobee	33102	91.1	7.5	14.8	14.6	59.1	9.8	15162
Orange	783974	79.1	17.5	12.3	10.4	78.8	21.2	20469
Osceola	142128	90.7	6.6	15.3	13.2	73.7	11.2	16256
Palm Beach	1018524	83.9	14.4	9.8	23.7	78.8	22.1	33518
Pasco	320253	96.5	2.3	4.4	32.0	66.9	9.1	16924
Pinellas	871766	89.1	9.0	3.1	26.6	78.1	18.5	24796
Polk	448646	83.3	15.4	5.3	18.2	68.0	12.9	17824
Putnam	70430	78.1	20.9	3.4	17.6	64.3	8.3	14250
Santa Rosa	114481	92.6	4.6	2.0	8.9	78.5	18.6	17127
Sarasota	301644	94.0	5.1	2.8	32.3	81.3	21.9	30205
Seminole	344729	87.4	9.8	8.4	10.1	84.6	26.3	21815
St. Johns	112707	88.7	10.1	3.0	15.6	79.9	23.6	25637
St. Lucie	179559	79.6	19.0	5.2	20.1	71.7	13.1	16483
Sumter	39428	81.0	18.1	3.1	20.3	64.3	7.8	14606
Suwannee	33077	82.2	16.9	2.0	15.8	63.8	8.2	14773
Taylor	18718	77.1	21.5	1.3	12.7	62.1	9.8	15459
Union	12359	71.0	27.8	4.8	7.0	67.7	7.9	10783
Volusia	419797	88.0	10.5	5.0	22.7	75.4	14.8	17778
Wakulla	19172	83.9	14.9	.9	10.9	71.6	10.9	15570
Walton	37914	88.9	8.6	1.2	14.9	66.5	11.9	14866
Washington	20221	79.7	17.6	1.5	16.4	60.9	7.4	13732

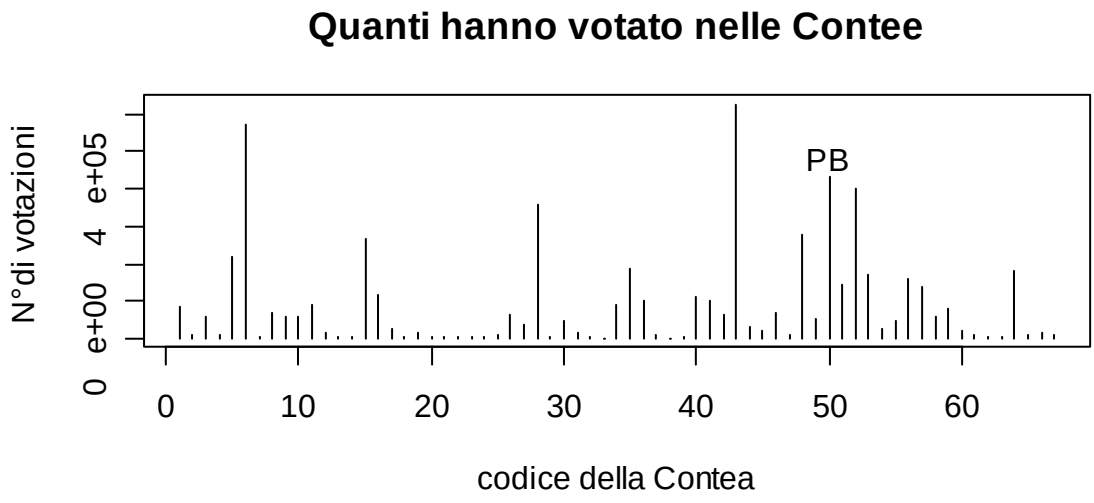
County	Bush	Gore	Brow	Nade	Har	Hag	Buc	Mc	Ph	Mo
Alachua	34124	47365	658	3226	6	42	263	4	20	21
Baker	5610	2392	17	53	0	3	73	0	3	3
Bay	38637	18850	171	828	5	18	248	3	18	27
Bradford	5414	3075	28	84	0	2	65	0	2	3
Brevard	115185	97318	643	4470	11	39	570	11	72	76
Broward	177323	386561	1212	7101	50	129	788	34	74	124
Calhoun	2873	2155	10	39	0	1	90	1	2	3
Charlotte	35426	29645	127	1462	6	15	182	3	18	12
Citrus	29765	25525	194	1379	5	16	270	0	18	28
Clay	41736	14632	204	562	1	14	186	3	6	9
Collier	60433	29918	185	1399	7	34	122	4	10	29
Columbia	10964	7047	127	258	1	7	89	2	8	5
Desoto	4256	3320	23	157	0	0	36	3	8	2
Dixie	2697	1826	32	75	0	2	29	0	3	2
Duval	152098	107864	952	2757	37	162	652	15	58	41
Escambia	73017	40943	296	1727	6	24	502	3	110	20
Flagler	12613	13897	60	435	1	4	83	3	3	12
Franklin	2454	2046	17	85	1	3	33	0	3	2
Gadsden	4767	9735	24	139	3	4	38	4	7	6
Gilchrist	3300	1910	52	97	0	1	29	0	2	4
Glades	1841	1442	12	56	0	3	9	1	0	1
Gulf	3550	2397	21	86	2	4	71	2	2	9
Hamilton	2146	1722	12	37	4	1	23	8	7	4
Hardee	3765	2339	17	75	0	2	30	0	2	3
Hendry	4747	3240	11	103	3	1	22	2	7	2
Hernando	30646	32644	116	1501	8	26	242	4	10	22
Highlands	20206	14167	64	545	6	16	127	3	7	8
Hillsborough	180760	169557	1138	7490	35	217	847	29	68	154
Holmes	5011	2177	18	94	1	7	76	3	6	2
Indian River	28635	19768	122	950	4	13	105	2	13	10
Jackson	9138	6868	40	138	0	2	102	1	4	7
Jefferson	2478	3041	14	76	2	1	29	1	0	0
Lafayette	1670	789	6	26	2	0	10	1	1	0
Lake	50010	36571	204	1460	4	36	289	1	21	15
Lee	106141	73560	538	3587	30	81	305	5	34	96
Leon	39053	61425	330	1932	9	28	282	7	16	31
Levy	6858	5398	92	284	1	1	67	1	10	12
Liberty	1317	1017	12	19	0	3	39	0	1	2

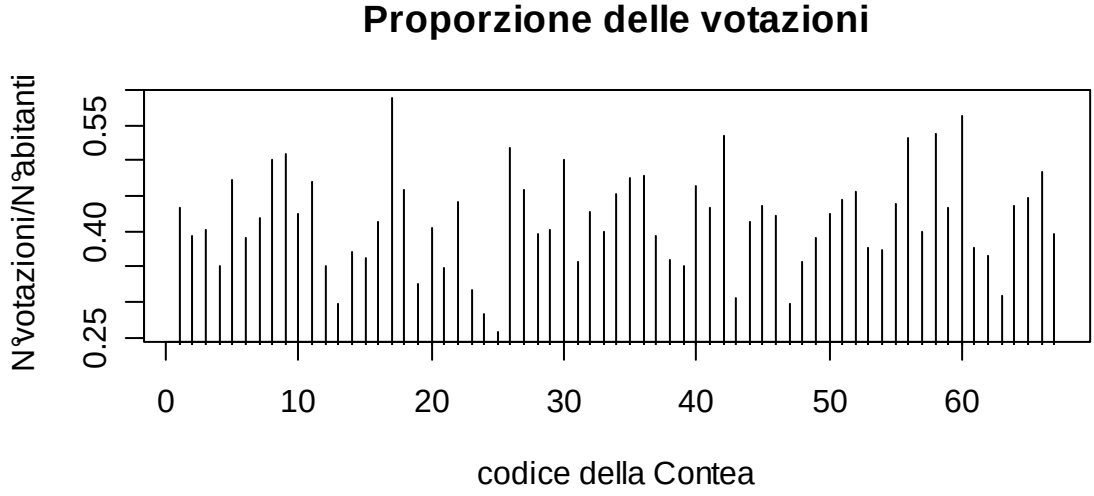
County	Bush	Gore	Brow	Nade	Har	Hag	Buc	Mc	Ph	Mo
Madison	3038	3014	18	54	0	2	29	1	1	5
Manatee	57952	49177	242	2491	5	35	271	3	19	26
Marion	55141	44665	662	1809	13	26	563	6	22	49
Martin	33970	26620	109	1118	14	29	112	7	20	14
Miami-Dade	289492	328764	760	5352	87	119	560	35	69	124
Monroe	16059	16483	162	1090	1	26	47	0	3	7
Nassau	16280	6879	62	253	0	7	90	4	3	3
Okaloosa	52093	16948	313	985	4	15	267	2	33	20
Okeechobee	5057	4588	21	131	1	4	43	1	3	4
Orange	134517	140220	891	3879	13	65	446	7	41	46
Osceola	26212	28181	309	732	10	20	145	5	10	33
Palm Beach	152846	268945	743	5564	45	143	3407	302	188	103
Pasco	68582	69564	413	3393	19	83	570	14	16	77
Pinellas	184823	200629	1230	10022	41	442	1013	27	72	170
Polk	90180	75193	365	2062	8	59	532	5	46	36
Putnam	13447	12102	114	377	2	7	148	3	10	12
Santa Rosa	36274	12802	131	724	1	13	311	1	43	19
Sarasota	83100	72853	431	4069	11	94	305	5	15	59
Seminole	75677	59174	550	1946	6	38	194	5	18	26
St. Johns	39546	19502	210	1217	4	11	229	2	12	13
St. Lucie	34705	41559	165	1368	4	12	124	10	13	29
Sumter	12127	9637	53	306	2	2	114	0	3	17
Suwannee	8006	4075	52	180	2	4	108	0	9	5
Taylor	4056	2649	4	59	0	3	27	1	8	1
Union	2332	1407	15	33	1	0	37	0	1	0
Volusia	82214	97063	442	2903	8	36	496	5	20	69
Wakulla	4512	3838	30	149	2	3	46	1	0	6
Walton	12182	5642	68	265	3	11	120	2	7	18
Washington	4994	2798	32	93	0	2	88	0	9	5

2.2 Analisi descrittiva delle variabili

Passiamo ora ad effettuare una prima analisi descrittiva delle variabili per avere una visuale ed un'idea migliore di come sono i dati che utilizzeremo in seguito per effettuare le nostre stime.





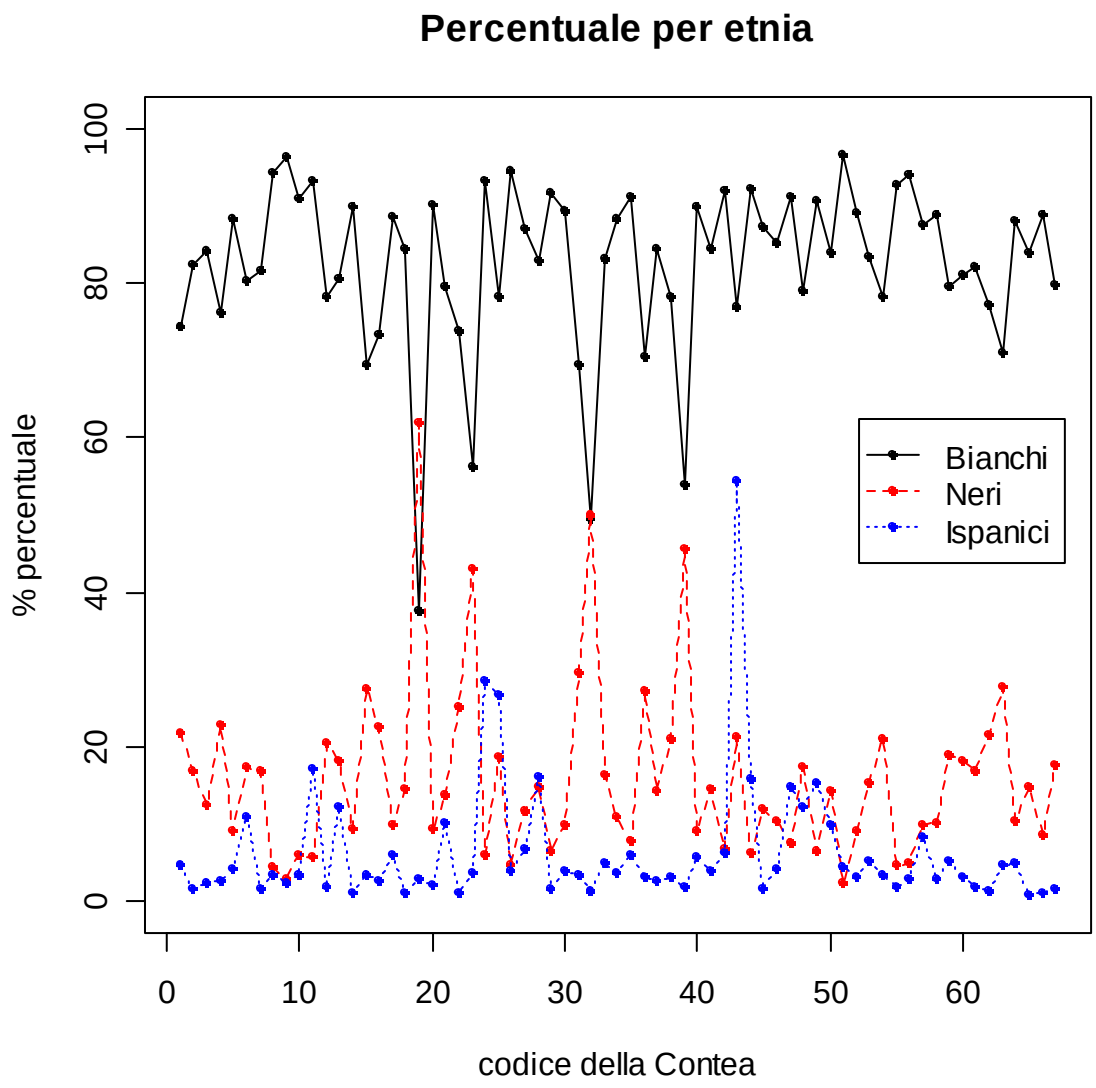


Nei tre grafici precedenti è stata analizzata la variabile della popolazione e la variabile, appositamente creata (“totv”), del numero totale di voti registrati. Dall’ultimo dei tre grafici notiamo che in tutta la Florida l’affluenza alle urne è stata abbastanza bassa, infatti, ha votato in media solo il 41% degli aventi diritto con un massimo del 59% nella contea di Flagler ed un minimo del 26% nella contea di Hendry. Per quanto riguarda la contea del capoluogo, Miami-Dade, ha riportato un 31% di votanti, mentre la contea di Palm Beach, soggetto di questo lavoro, ha registrato il 42% di schede valide.

Nei primi due grafici troviamo conferma di quanto detto nell’introduzione, vale a dire che la contea di Palm Beach è la terza sia in ordine di numero di abitanti, sia nel numero finale di schede votate. Inoltre, iniziamo a notare (graficamente è più facile che osservando i dati e basta) che la differenza di numerosità tra una contea e l’altra è sicuramente molto ampia e questo

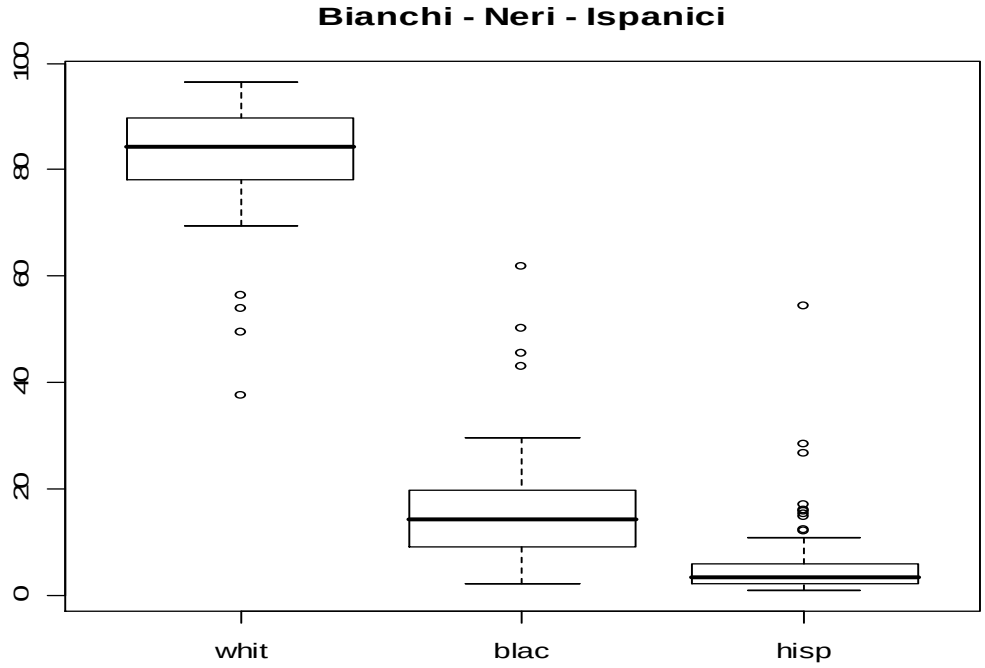
influenzerà per certo le analisi e le stime che faremo nel seguito di questo studio. Per esempio, notiamo che la differenza tra la contea con maggiore popolazione (Miami-Dade con 2.044.600 abitanti) e quella con minore (Lafayette con poco più di 6mila unità) è di più di 2 milioni di cittadini, questo non potrà che influenzare sul peso che ogni covariata avrà al cambio di contea su un modello stimato.

Passiamo ora a guardare le tre variabili che identificano la percentuale relative alla razza: Bianchi, Neri ed Ispanici. Ricordiamo inoltre che la somma dei tre in più di qualche contea supera il 100% visto che in “hisp” sono state registrate più etnie (ad esempio in Miami-Dade la somma supera il 150%).



Questo grafico, i seguenti boxplot, il diagramma a torta e gli istogrammi mostrano semplicemente la distribuzione delle varie razze nelle contee della Florida. Le sole osservazioni che possiamo fare riguardano la netta

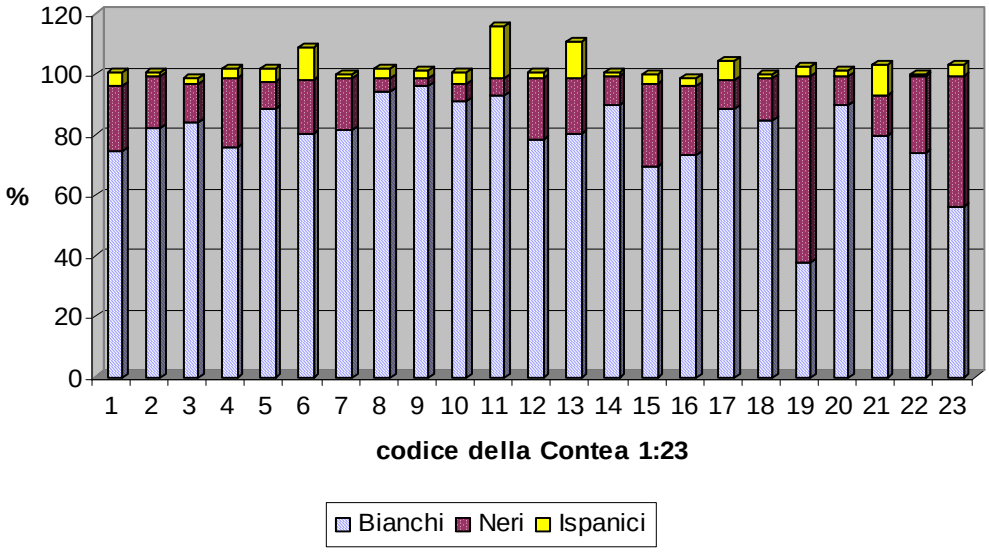
maggioranza di bianchi in tutte le frazioni e, tranne che per Miami che registra un picco del 54,4%, una scarsa presenza d'ispanici.

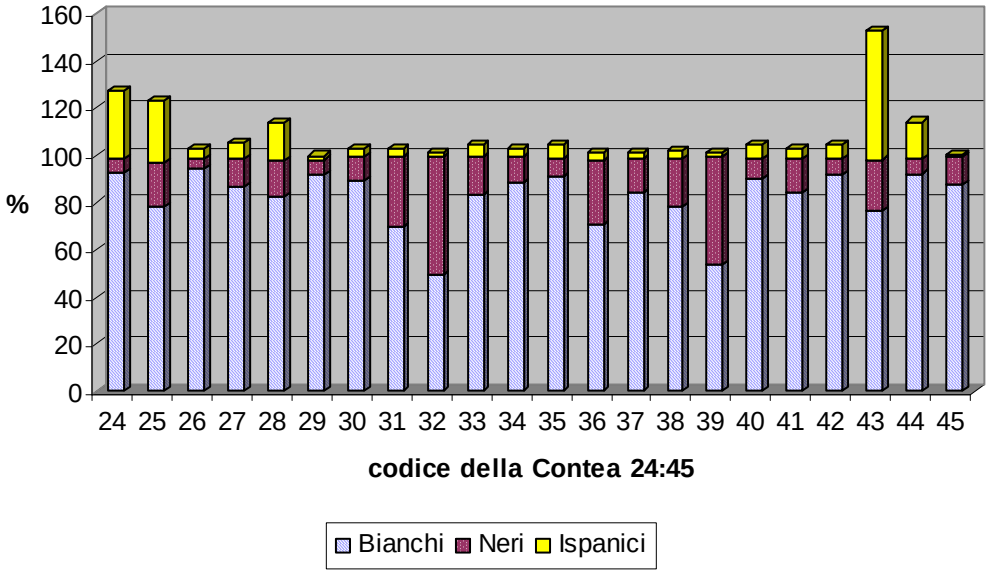


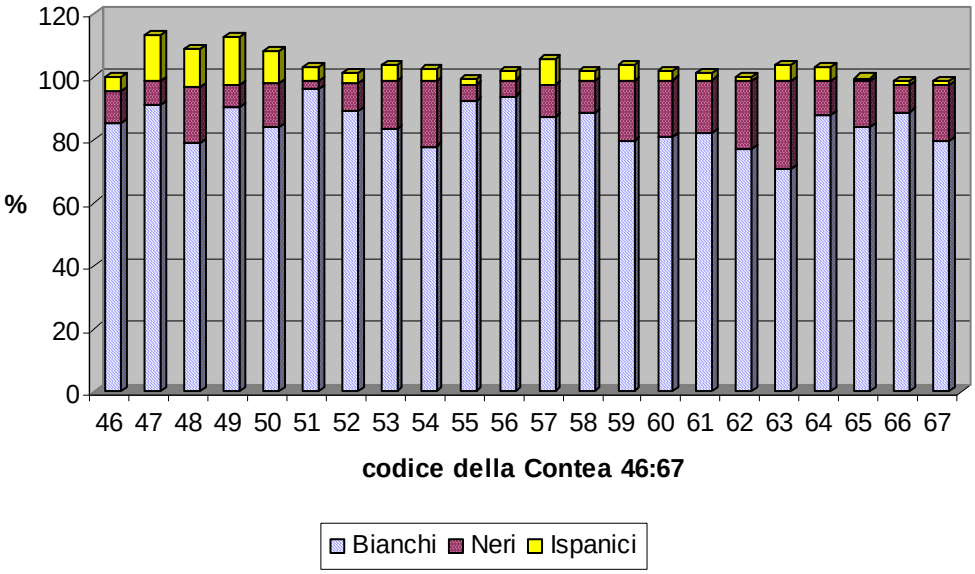
	Bianchi	Neri	Ispanici
Minimo	37.60	2.30	0.90
1Quartile	78.25	9.00	2.25
Mediana	84.20	14.40	3.50
Media	82.42	15.90	6.29
3Quartile	89.80	19.75	5.90
Massimo	96.50	61.80	54.40

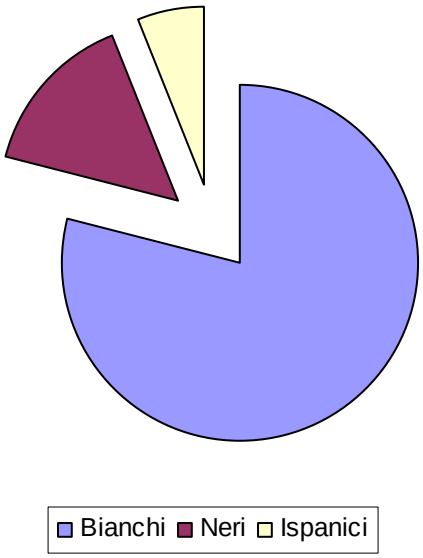
Tutti i valori della tabella sono espressi in percentuale %

La contea al centro del nostro studio, Palm Beach, rientra pienamente nell'andamento medio dello stato, infatti registra 83,9% di bianchi, il 14,4% di neri e il 9,8% d'ispanici

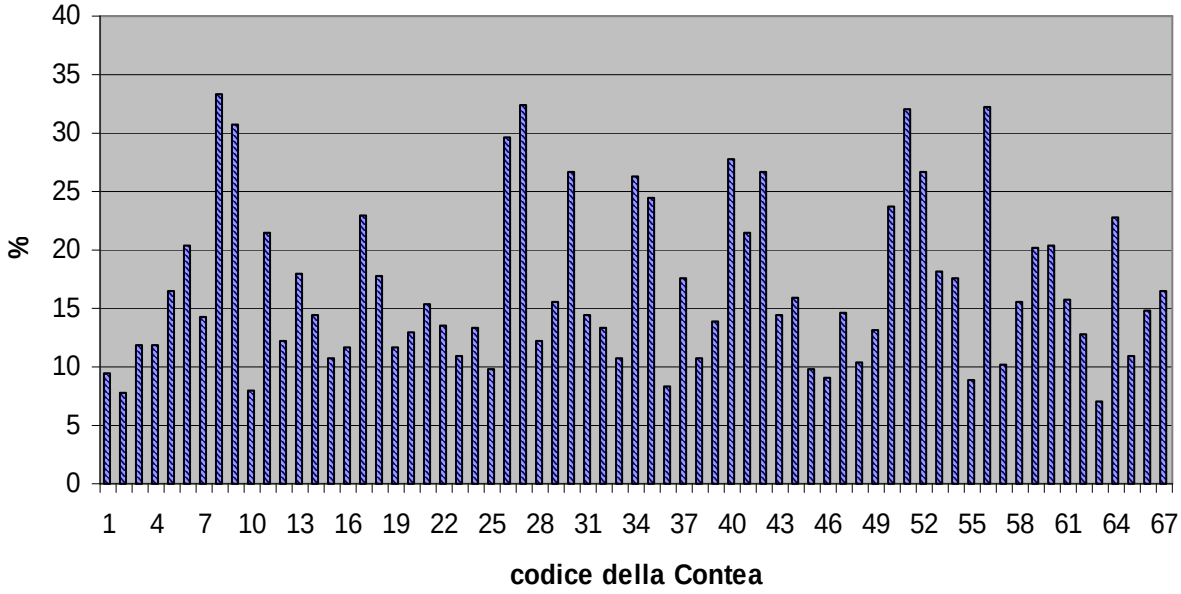

120
100
80
%
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
codice della Contea 1:23
Bianchi Neri Ispanici


160
140
120
100
%
80
60
40
20
0
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
codice della Contea 24:45
Bianchi Neri Ispanici


120
100
80
%
60
40
20
0
46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67
codice della Contea 46:67
Bianchi Neri Ispanici



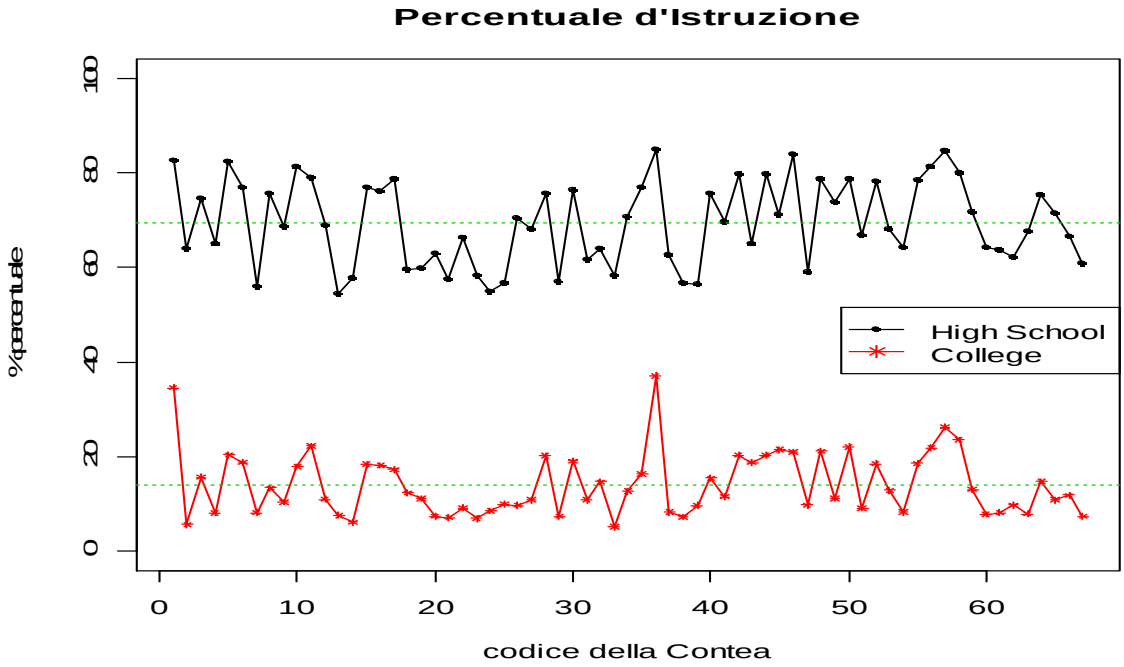
Segue l'istogramma della variabile ≥65 in cui viene visualizzata la percentuale di persone (calcolata sulla popolazione totale d'ogni contea) con età superiore od uguale ai 65 anni:



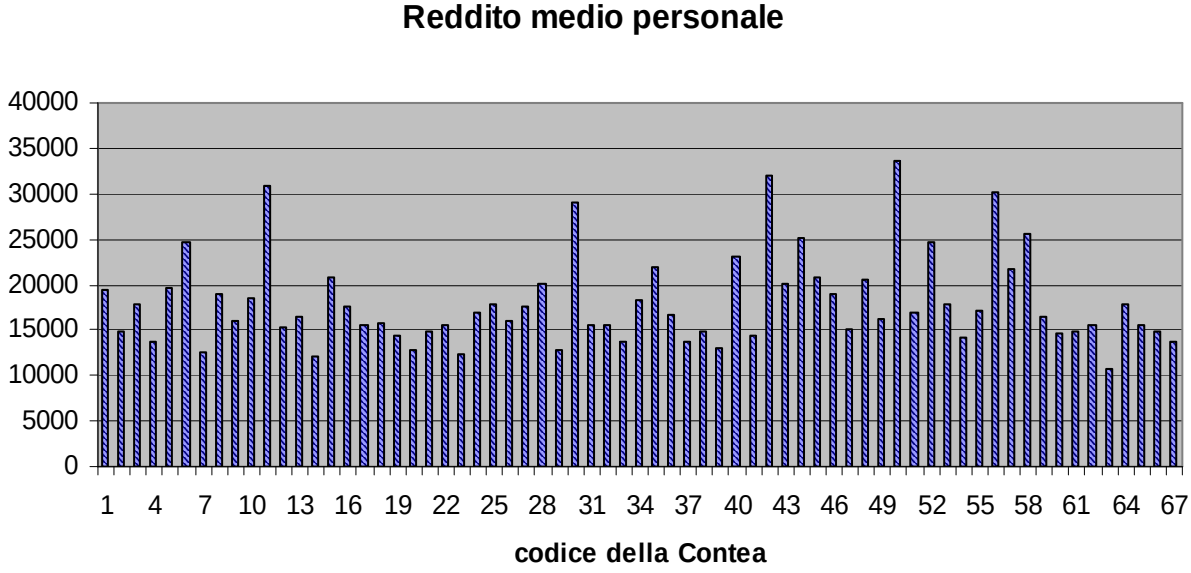
Palm Beach, in questo caso, si pone poco sopra la media di 16,8%, con una percentuale di over 65enni del 23,7%

Nel grafico successivo sono rappresentate le percentuali relative al numero di persone che, per ogni contea, hanno concluso l'high school o il college. E' stata aggiunta, per ciascun andamento, la linea che rappresenta la media per ogni livello d'istruzione, che risulta essere per HS: 69.5% mentre per il College: 14%. Come si può notare, i valori di Palm Beach emergono

dall'andamento medio registrando, in entrambi i casi,un livello d'istruzione maggiore a quello della media statale, difatti, secondo i dati forniti dal Census, i cittadini della contea di PB risultano aver superato le scuole superiori per il 78,8% e aver terminato il college il 22,1%.



Per ultima, tra le variabili demografiche, abbiamo l'indicatore del reddito medio personale. Dall'istogramma possiamo notare che la contea col valore medio di reddito più alto è la 50^{a} , ovvero Palm Beach che risulta avere un valore di oltre una volta e mezzo maggiore rispetto alla media dello stato della Florida.



Traendo una veloce conclusione dall'analisi delle variabili demografiche fatta in queste pagine, risulta che Palm Beach, si posiziona tra le contee più importanti dell'intera Florida. Compare, infatti, che oltre ad avere una popolazione numerosa e un numero di schede registrate tra i più alti (è la terza in ordine decrescente), si registra un alto livello d'istruzione e una ricchezza media molto elevata (sempre in rapporto con la media dello stato). Questo fa pesare maggiormente l'ipotesi di errore di voto dichiarata in questa contea e alla "nascita improvvisa" di una fede riformista (i 3407 voti ottenuti da Buchanan) mai registrata prima in questa zona da sempre democratica.

Passiamo ora ad esaminare le variabili relative agli esiti elettorali; trascureremo, per adesso, la variabile soggetto del nostro studio, "buch" (relativa ai voti ottenuti da Buchanan, il candidato del partito riformista) che riprenderemo più attentamente in seguito.

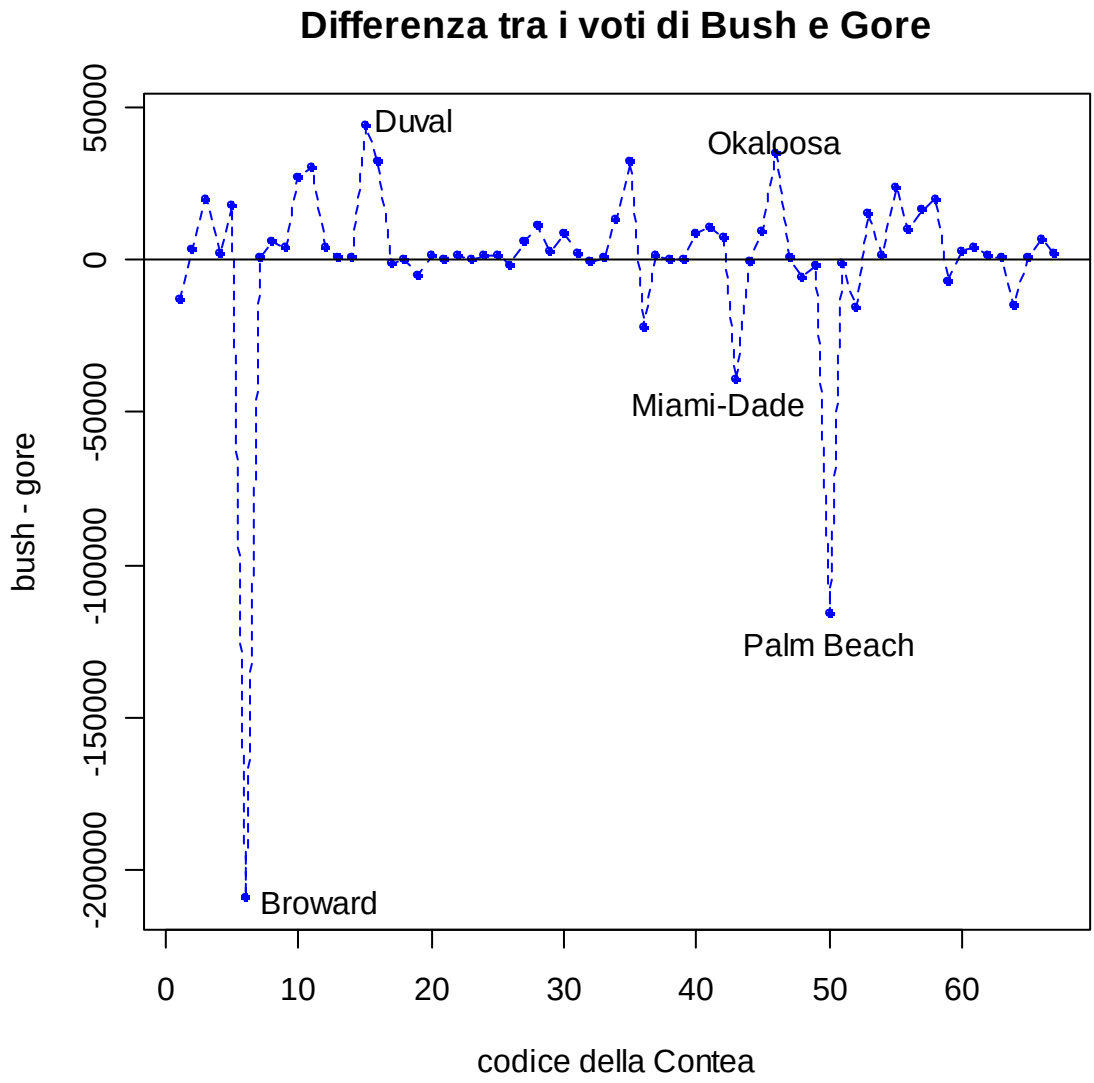
Partiamo con un po' di valori:

	N° Totale di voti	% Complessiva
Bush	2910078	48,85%
Gore	2909117	48,83%
Browne	16396	0,26%
Nader	97416	1,63%
Harris	558	0,00009%
Hagelin	2273	0,0004%
Buchanan	17465	0,003%
McReynolds	618	0,0001%
Phillips	1368	0,0002%
Moorehead	1803	0,0003%

Già si capisce, osservando i dati riportati nella tabella, che i risultati ottenuti dagli ultimi 6 candidati sono poco influenti sugli altri e non saranno rilevanti al fine della stima del modello. Infatti, se Bush e Gore assieme hanno ottenuto quasi il 98% dei voti, del restante 2%, l'1,99% è andato a Nader e Browne e solo lo 0,01% agli altri sei.

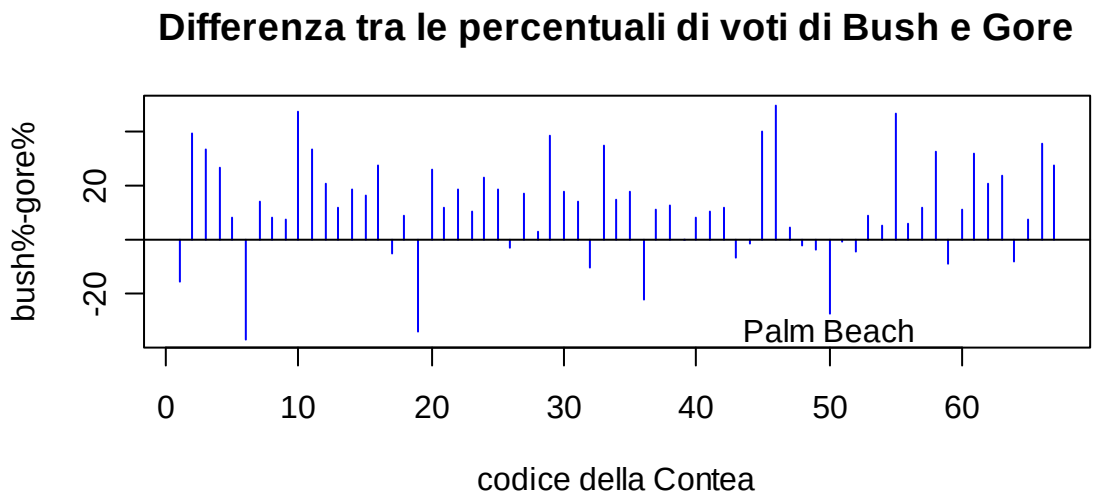
Andiamo ora ad osservare il rapporto che c'è tra i voti di Bush e quelli di Gore utilizzando tre grafici: il primo relativo alla differenza tra i voti dell'uno e quelli dell'altro, il secondo relativo alla differenza delle percentuali e il terzo è un

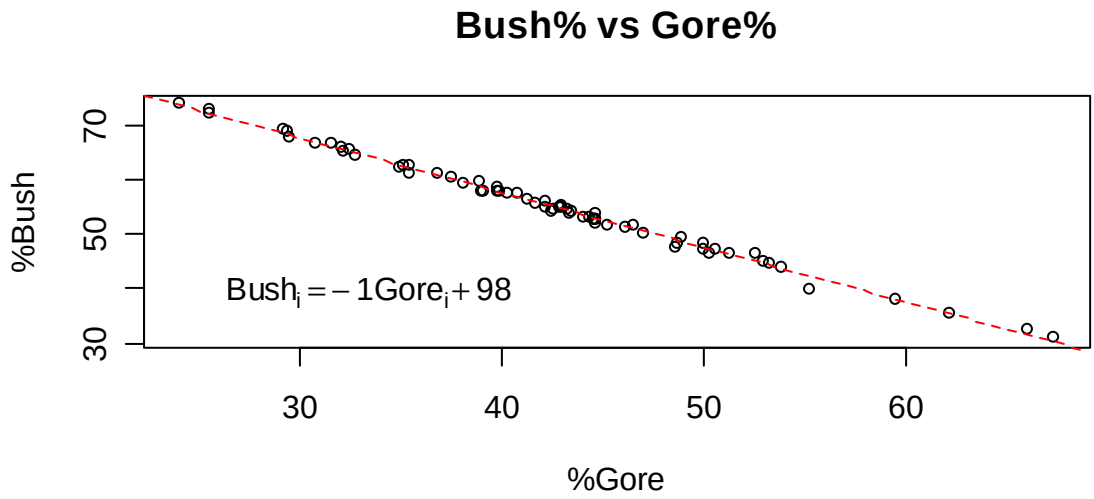
grafico di dispersione tra le percentuali di voti ottenuti dai due candidati al quale è stata aggiunta la retta dell'andamento, stimata tramite il comando in R "lm", che genera un modello di regressione lineare normale e con il quale possiamo disegnare la retta di regressione.



Avendo calcolato i valori dell'espressione Bush - Gore, nel grafico sopra riportato sono rappresentati in positivo le contee vinte da Bush, ed in negativo quelle ottenute da Gore. Notiamo che il candidato repubblicano ha vinto nel maggior numero di contee, anche se di pochi voti, mentre il democratico Gore ha registrato dei picchi in alcune contee; in particolare si nota, come detto nell'introduzione, la forte tendenza di Palm Beach verso una fede democratica e ad un'adesione al candidato Gore. In essa, difatti, il partito democratico ha ricevuto il 62,2% di assensi contro solo il 35,3% di voti per i repubblicani.
Il grafico successivo ci conferma la vittoria di Bush nella maggior parte delle

Contee, con scarti spesso superiori al 20%, ma contando per la vittoria finale il numero totale di voti in tutta la Florida, le vittorie di Gore, si piazzarono nelle contee più popolate (idea ben espressa dal grafico precedente) rendendo la differenza finale minima.



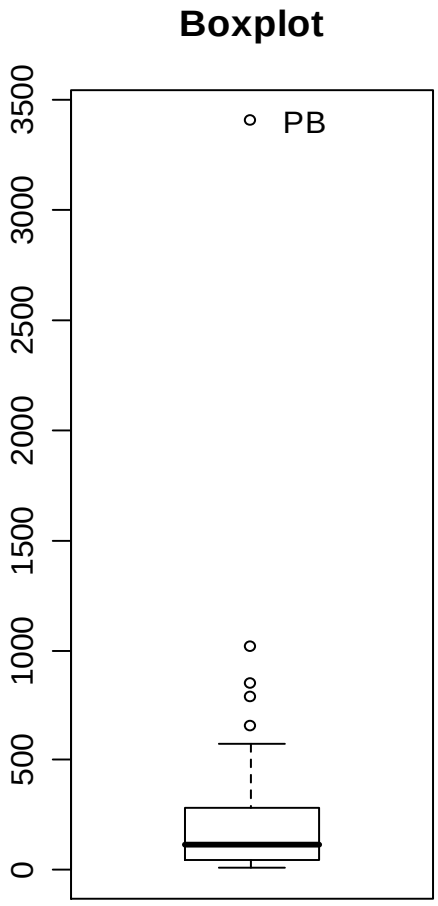


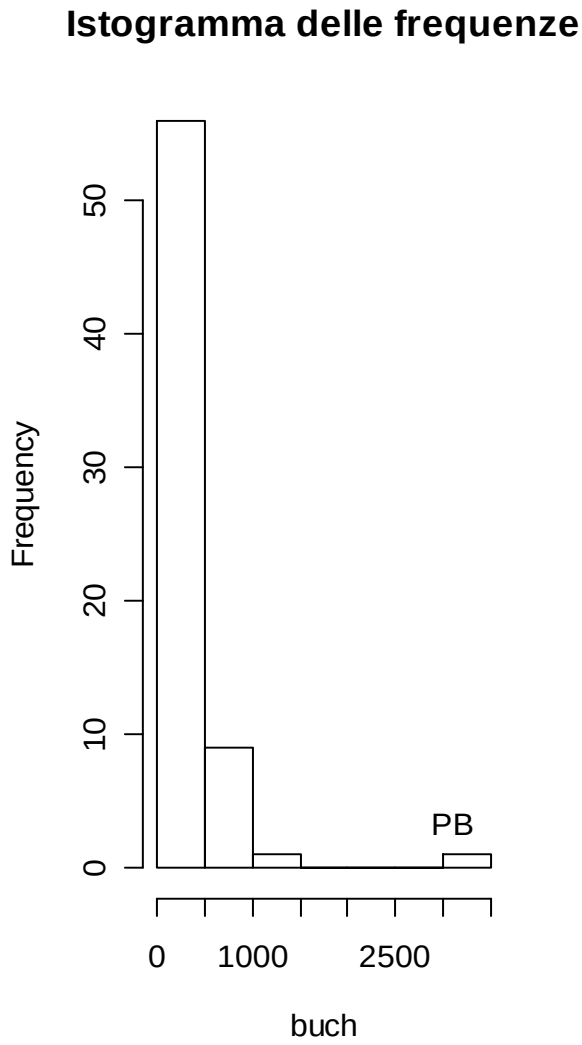
Da questo secondo grafico si capisce subito la linearità del rapporto tra la percentuale di voti raccolti da Bush e quella ottenuta da Gore. La retta attorno alla quale si distribuiscono perfettamente i dati ha espressione: $Bush\% = -Gore\% + 98\%$ e ci indica che le due variabili sono decisamente inversamente proporzionali, ed al crescere di una diminuisce l'altra, difatti, a basse percentuali di Gore, il risultato ottenuto da Bush si avvicinerà al 100% e, contrariamente, a scarsi risultati per il repubblicano corrisponderanno alte percentuali di voti per il democratico. Questo ci fa affermare che quando andremo a stimare un modello lineare, nei prossimi capitoli, potremo omettere

una delle due variabili, tanto sappiamo che una avrà coefficiente stimato opposto all'altra.

2.3 Analisi della variabile "buch"

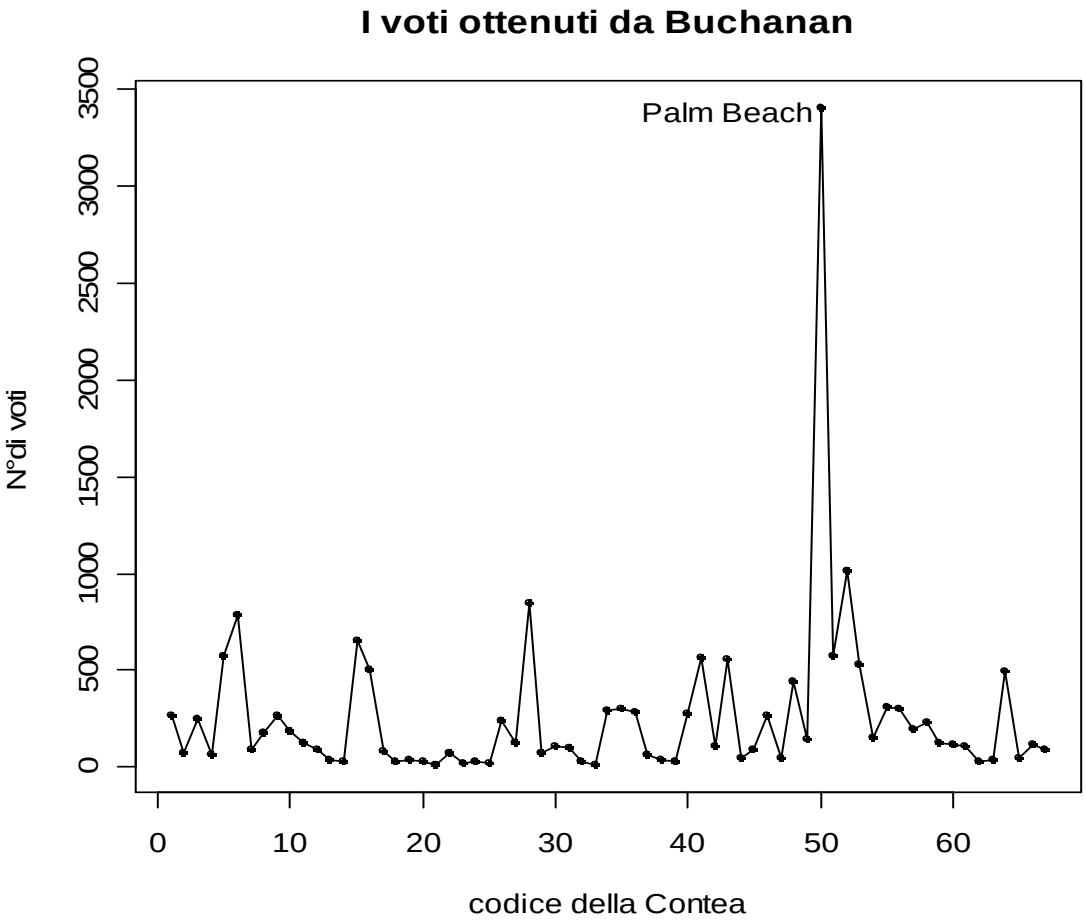
Andiamo adesso a studiare la variabile relativa ai voti ottenuti dal candidato Pat Buchanan del partito riformista, ponendo particolare attenzione ai risultati registrati nella contea di Palm Beach, e alla relazione con le altre variabili per farci un'idea su quali saranno significative o quali eventuali trasformazioni potranno esserci utili.

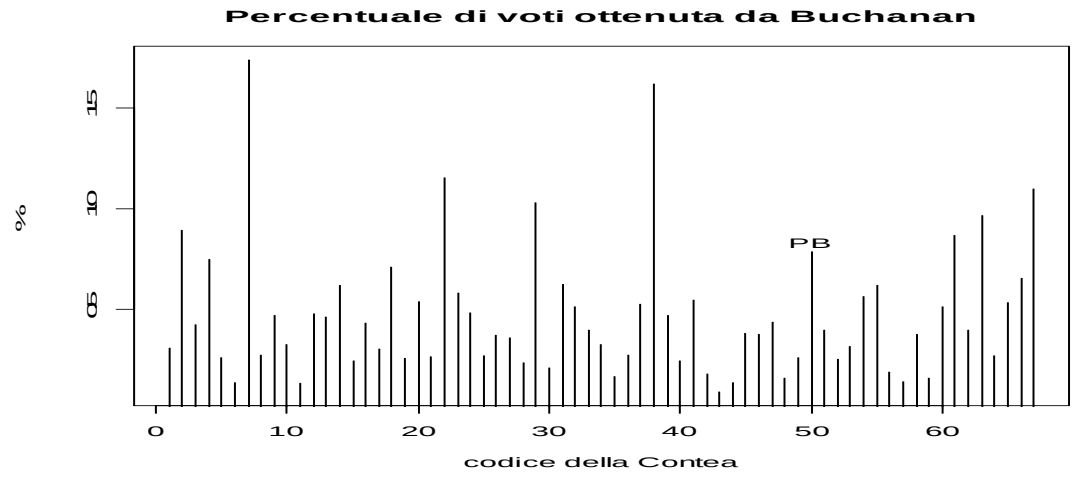




Min	Max	Media	1°Quart	3°Quart	Mediana
9,0	3407	260,7	46,5	285,5	120

Dal boxplot e dall'istogramma sulle frequenze della variabile "buch", notiamo chiaramente due outliers (identificati dalla sigla PB) che rappresentano entrambi il valore relativo alla contea di Palm Beach. Essi rappresentano casi separati dalla media degli altri risultati, basti pensare che il risultato migliore ottenuto da Buchanan, dopo Palm Beach, è quello registrato a Pinellas con appena 1013 voti, ben 2034 voti di differenza, e che la media dei voti è di appena 260,7 per contea.

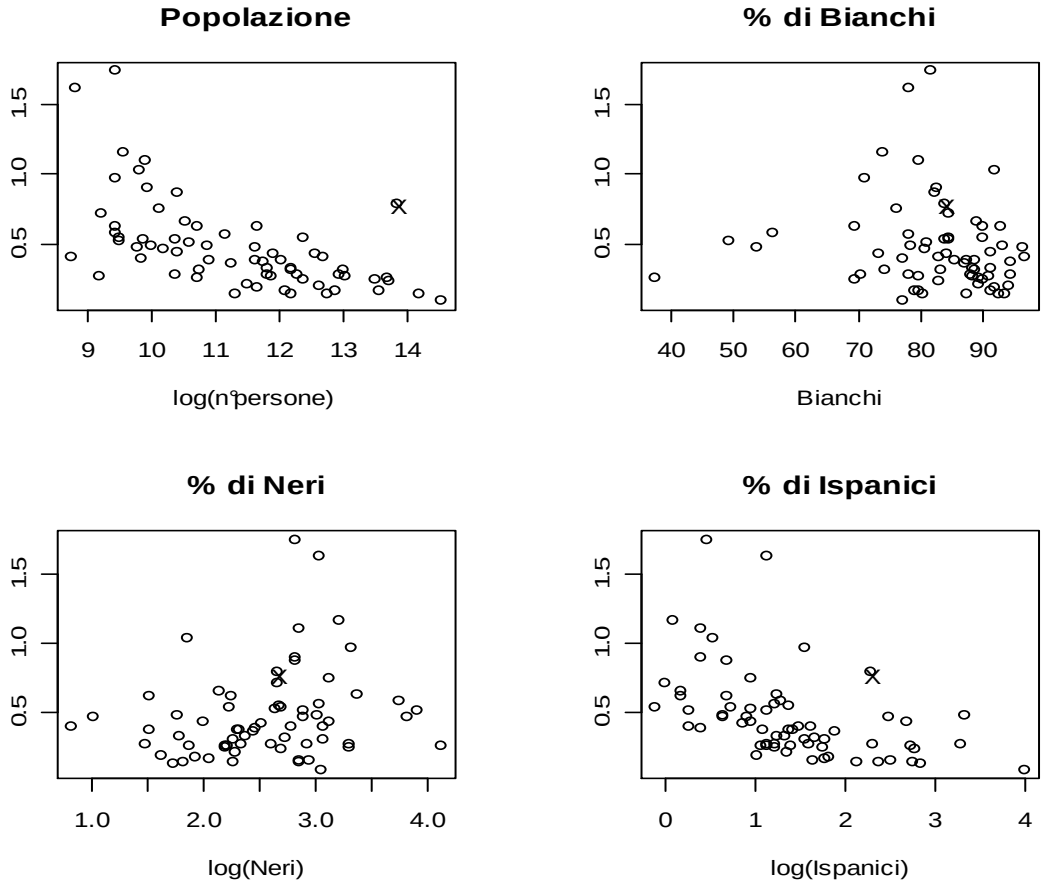


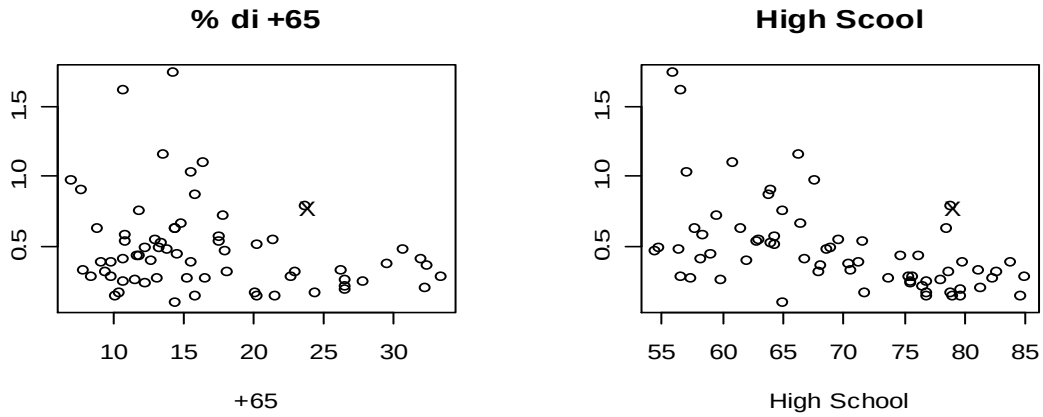


Osservando anche questi due grafici, si nota subito la grande differenza di voti tra quelli registrati a Palm Beach e gli altri. Escludendo, ad esempio, la 50^{a} osservazione dal calcolo della media, il valore ottenuto tra le altre 66 contee è di 213 voti, mentre, includendo nuovamente la contea di PB, il valore della media cresce di ben 47,7 voti, E' chiaro che questo non fa altro che confermare i nostri sospetti sull'erroneità del valore ottenuto e sul corretto svolgimento delle elezioni.

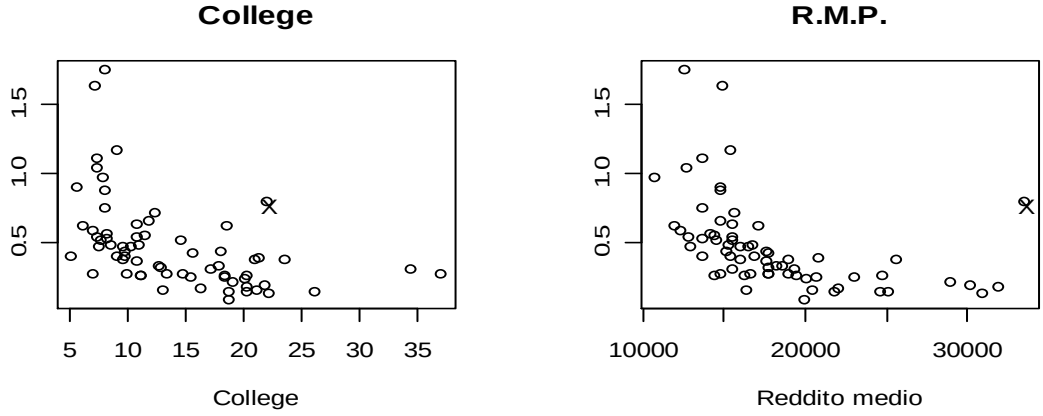
Passiamo ora ad osservare il rapporto tra "buch" e le altre variabili, ponendo sempre attenzione alla 50^{a} osservazione (quella relativa alla contea di Palm Beach). Useremo per tutte le variabili relative agli esiti elettorali, al posto del numero totale di voti, quello relativo alla percentuale di schede ottenute ed applicheremo ad alcune variabili demografiche la trasformazione logaritmica per ottenere un responso più soddisfacente.

Nei grafici che seguono, sull'asse delle ordinate c'è la variabile riguardante la percentuale di voti complessiva ottenuta da Buchanan, e l'osservazione corrispondente al dato ottenuto nella contea di Palm Beach è stata contraddistinta da una X.

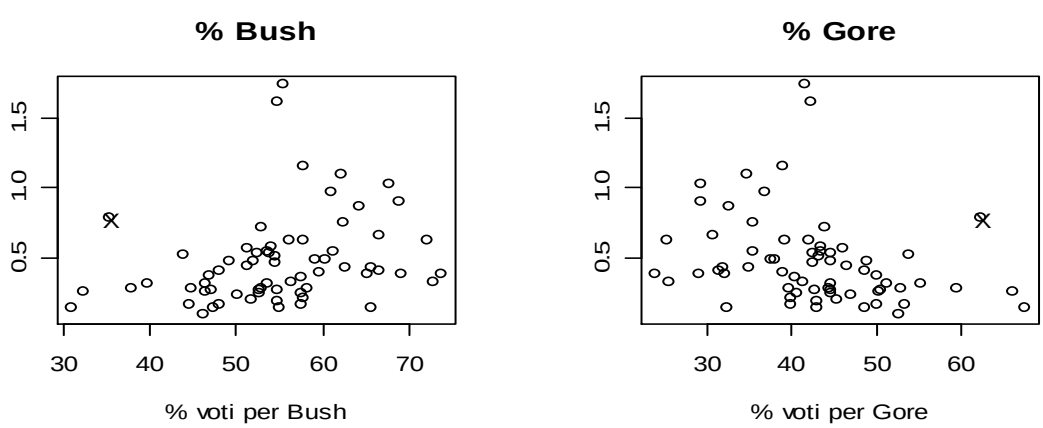



% di +65
0.5
1.0
1.5
10
15
20
25
30
+65

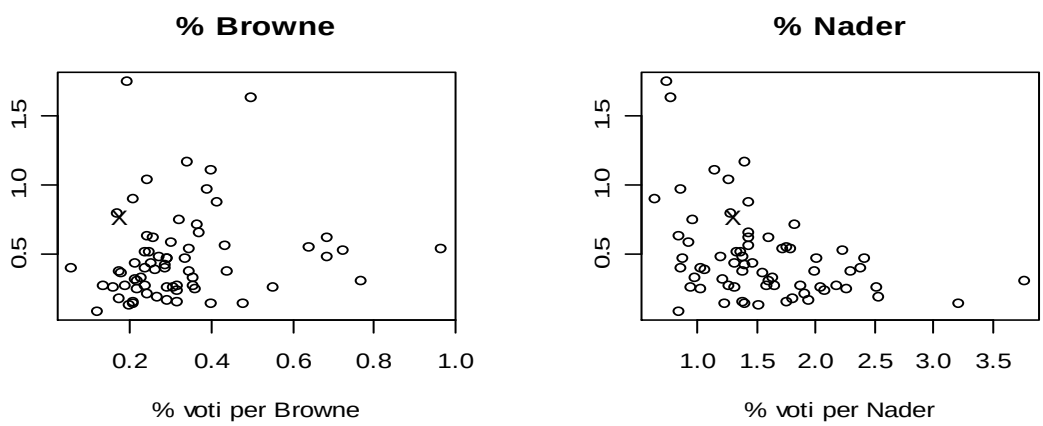
High Scool
0.5
1.0
1.5
55
60
65
70
75
80
85
High School


College
0.5
1.0
1.5
5
10
15
20
25
30
35
College

R.M.P.
0.5
1.0
1.5
10000
20000
30000
Reddito medio


% Bush
0.5
1.0
1.5
30
40
50
60
70
% voti per Bush

% Gore
0.5
1.0
1.5
30
40
50
60
% voti per Gore


% Browne
0.5
1.0
1.5
0.2
0.4
0.6
0.8
1.0
% voti per Browne

% Nader
0.5
1.0
1.5
1.0
1.5
2.0
2.5
3.0
3.5
% voti per Nader

Le variabili a cui è stata applicata la trasformazione logaritmica sono “npop”, “blac” e “hisp”, ovvero quelle relative alla popolazione, alla percentuale di neri e alla percentuale d’ispanici. I grafici relativi al rapporto tra Buchanan e la percentuale degli altri candidati non sono stati riportati perché ininfluenti ai fini del nostro studio, a causa del basso numero di voti ottenuti; difatti, con meno dello 0,001% complessivo di schede a favore, la presenza o meno di queste variabili nello studio è indifferente. Includerle vorrebbe dire tentare di dare un peso alle poche migliaia (per l’esattezza circa 6600) di voti ottenute dagli altri candidati, cosa alquanto irrilevante visto che stiamo parlando di un complessivo di voti che si aggira attorno ai 6 milioni.

3. STIMA DEL MODELLO

Dopo aver analizzato le singole variabili ed aver osservato le relazioni tra “buch” e tutti gli altri fattori, passiamo a stimare un adeguato modello statistico per poter, in seguito, cercare di fare delle previsioni su quanti voti avrebbe dovuto ottenere Buchanan nella contea di Palm Beach.

3.1 Le variabili impiegate

Per poter stimare un modello adeguato alle nostre esigenze ed ai nostri scopi dobbiamo apportare alcune modifiche al data set finora utilizzato. Infatti, lo scopo del nostro studio è quello di arrivare a predire il numero di voti che il candidato riformista avrebbe dovuto ottenere nella zona di Palm Beach basandoci sui risultati ottenuti nelle altre 66 contee; per fare questo è importante sostituire il risultato originale (3407 voti), ritenuto da noi un errore e che quindi potrebbe influenzare negativamente la stima dei parametri del modello, e sostituirlo con un valore mancante. Per questo, nel file utilizzato d’ora in poi, il 50°valore della variabile “buch” (relativo ai voti ottenuti da Buchanan a Palm Beach) è stato cambiato con “NA”, che in R identifica, appunto, un valore mancante.

Da questo passaggio ne deriva un altro, infatti, per poter fare delle buone previsioni, dobbiamo, a questo punto, ricalcolare anche il valore di N_i , ovvero il numero totale di voti registrati, eliminando quelli ottenuti da Buchanan dalla

somma, superando, così, il problema del dato mancante. Non possiamo, infatti, continuare a tenere la vecchia variabile avendo posto come erratica l'osservazione ricavata nella contea di Palm Beach. Questa operazione non crea ai calcoli ed alle analisi successive grandi difficoltà e non ne diminuisce il valore e la correttezza in maniera rilevante, difatti la percentuale di voti omessa (quelli del candidato riformista) dalla somma non supera lo 0,003% dei voti totali influenzando, di conseguenza, solo marginalmente i valori finali.

L'importanza della riparametrizzazione di N_i è relativa anche al fatto che questa variabile verrà utilizzata, in seguito, per scalare tutte le covariate dei modelli che andremo a calcolare. Come abbiamo potuto osservare, infatti, dai grafici ottenuti analizzando la variabile sulla popolazione o quelli relativi al numero totale di voti registrati, era facile notare la grande differenza di valori tra una contea e un'altra. Questo ci porta a considerare che non sia pensabile che l'influenza sulla variabile risposta di una qualsiasi covariata, come il reddito medio ad esempio, non sia differente tra la contea di Miami-Dade, con 2044600 abitanti, e la contea di Lafayette, con soli 6289 abitanti. Logicamente, infatti, il livello dell'effetto della variabile "inco" sui voti ottenuti da un candidato, sarà relativo alla probabilità che un singolo votante supporti tale candidato. Di conseguenza se si vuole sapere come il reddito influenza e pesa sulla variabile risposta basterà moltiplicare il valore del reddito per il numero totale di voti registrati nella relativa contea. Questo ragionamento vale per tutte le covariate che di conseguenza verranno riscalate, ovvero verranno moltiplicate per il numero totale di voti N_i (ricalcolato come detto in precedenza, quindi con l'omissione dei voti ottenuti da Buchanan).

Quindi, essendo y_i la variabile risposta, ed $x_{ij}\,\beta_j$ la combinazione lineare tra una qualsiasi covariata e il suo coefficiente stimato nel modello(con $i = 1,....,67$ il numero-codice della contea e $j = 1,...., p$ il numero di covariata), per le previsioni finali useremo $\left(x_{ij}\beta_j\right)^* = N_i \cdot \left(x_{ij}\beta_j\right)$. Essendo, però, N_i costante per ogni contea i, otterremo $N_i \cdot \sum_{j=1} x_{ij}\beta_j$ dove, di conseguenza, avremo che $N_i \cdot \eta = \eta^*$ con η predittore lineare del modello (concetto che vedremo meglio in seguito). Notiamo, dunque, che l'effetto di N_i non influenza la stima dei

parametri β_j, ma comporta una trasformazione dei valori solo dopo che il modello è già stato stimato.

Proprio per questo motivo di rivalutazione dei parametri è importante che il valore di N_i relativo al numero di votanti per ogni contea i, venga ricalcolato in base alla modifica del data set e all'eliminazione del dato outlier di Buchanan a Palm Beach, eliminando dalla somma tutti i voti ottenuti dal candidato riformista.

Segue, ora, la lista di tutte le variabili che utilizzeremo per stimare i modelli:

Covariate	Spiegazione
lnpop	Logaritmo della percentuale della popolazione
whit	Percentuale di bianchi
lblac	Logaritmo della percentuale di neri
lhisp	Logaritmo della percentuale d'ispanici
o65	Percentuale di persone con età ≥ 65
hsed	Percentuale di persone che hanno concluso l'High School
coll	Percentuale di persone che hanno concluso il College
incot	Reddito medio/1000 (modificato così per facilità d'interpretazione)
pbush	Percentuale di voti ottenuti da Bush
pbrow	Percentuale di voti ottenuti da Browne
pnade	Percentuale di voti ottenuti da Nader

Le tre trasformazioni logaritmiche sono suggerite dall'analisi descrittiva fatta in precedenza, infatti la popolazione, la percentuale di neri e la percentuale di ispanici fornivano un responso più soddisfacente scalate in questa maniera piuttosto che lasciate nella forma originale.

Non verranno utilizzate le percentuali di voto ottenute dai candidati Gore, Harris, Hagelin, McReynolds, Phillips, Moorehead. Il primo, Gore, perché abbiamo dimostrato in precedenza la forte correlazione con i voti ottenuti dal candidato Bush, quindi sarebbe una covariata totalmente inutile in un modello visto che ad alti valori di una corrisponderebbero bassi dell'altra e viceversa (si otterrebbero, in ogni caso, risultati molto simili a quelli che vedremo sostituendo ai voti di Bush quelli di Gore e usando lo stesso coefficiente di regressione con segno opposto).

Per i restanti candidati, l'esclusione dalla stima del modello non è dovuta a bassa correlazione con i voti di Buchanan o altro, ma alla bassissima percentuale di preferenze ottenute, il risultato più alto lo ha ottenuto Hagelin con lo 0,0004% di schede a favore; dati che ci portano ad omettere senza problemi le variabili elencate come già spiegato in precedenza.

3.2 Introduzione ai modelli lineari generalizzati

Prima di passare direttamente ai modelli generalizzati, guardiamo cos'è un modello di regressione lineare normale; esso, infatti, è una relazione che lega ad una variabile un insieme di covariate linearmente indipendenti espressa come:

VARIABILE RISPOSTA = f (VARIABILI ESPLICATIVE) + errore

Questi modelli (come quelli generalizzati) sono utili per lo studio delle relazioni tra dati raccolti su più variabili prese da uno stesso gruppo di unità statistiche. Avendo, ad esempio, y_i variabile risposta (o dipendente) e un insieme di covariate x_{ij} (variabili esplicative o indipendenti) , e posto che le osservazioni y_i siano una realizzazione della variabile casuale Y_i, dove $Y_i \sim N(\mu_i, \sigma^2)$ con $\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ (in forma matriciale $\mu_i = X\beta$ con X matrice delle covariate x_{ij}, e di conseguenza $Y_i \sim N_n(X\beta, \sigma^2)$), avremo che la relazione $Y_i = \mu_i + \varepsilon$ oppure $Y_i = X\beta + \varepsilon$ sarà detta modello di regressione lineare normale, con $\varepsilon \sim N(0, \sigma^2 I_n)$.

In questo tipo di modello riusciamo quindi ad identificare tre componenti principali:

- Componente casuale $Y_i \sim N(\mu_i, \sigma^2)$
- Componente sistematica $\eta_i = X\beta$ con η_i detto predittore lineare, con $\eta_i = \mu_i$
- Componente erratica (stocastica) $\varepsilon \sim N(0, \sigma^2 I_n)$

L'impiego di questi modelli, però, è limitato. Essi infatti non riescono ad essere applicati a tutti i tipi di dati, ad esempio a variabili risposta dicotomiche, qualitative o discrete; oppure può essere che Y_i non si distribuisca come una Normale ma piuttosto con una distribuzione discreta (Poisson, Gamma o

Binomiale); oppure può verificarsi che il legame tra il predittore lineare η_i e μ_i non sia lineare.
Per risolvere queste e altre possibili complicanze dei modelli lineari normali ed avere così stime e modelli più adeguati ai dati che si stanno adoperando ed analizzando, si utilizzano i modelli lineari generalizzati (in R sono i GLM).

I modelli lineari generalizzati sono un ampliamento dei modelli visti sopra (infatti li comprendono) e servono sempre per spiegare il rapporto tra una variabile risposta e più variabili esplicative, ma, a differenza di quanto osservato prima, sono maggiormente adattabili ai dati.
Come abbiamo potuto studiare in precedenza, le componenti dei modelli lineari normali erano tre; di queste, nei modelli generalizzati, ne rimangono due: infatti non sarà più possibile tenere separata la parte erratica ε da quella sistematica; in compenso però, si porrà maggiore attenzione al legame, prima solamente lineare, tra il predittore lineare η_i e il valor medio μ_i.

- Componente casuale: la nostra Y_i non dovrà più distribuirsi solamente come una normale, ma potrà spaziare tra tutte le distribuzioni della famiglia esponenziale (mantenendo l'ipotesi d'indipendenza). Questa famiglia molto ampia comprende distribuzioni come quella di Poisson, binomiale, normale, gamma, esponenziale, ecc. e le sue procedure d'inferenza per la stima dei parametri sono basate sulla verosimiglianza. La funzione di densità generale (che varierà da distribuzione a distribuzione al variare dei parametri) è:

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

con θ: parametro naturale della famiglia esponenziale

ϕ: parametro di dispersione o di scala (normalmente è fissato)

$b(\ ), c(\ )$: funzioni note che cambiano con le diverse distribuzioni

Da questa formula ne deriva, quindi, che $Y \sim EF(b(\theta),\phi)$ (EF :exponential family) con $E(Y) = b'(\theta) = \mu$ ed $\text{var}(Y) = b''(\theta) = \phi V(\mu)$

- Componente sistematica: questa parte non cambia, η_i rimane il predittore lineare e l'effetto delle covariate resta lineare $\eta_i = \sum_{j=1} x_{ij}\beta_i = X\beta$

- Legame: per definire, infine, un adeguato GLM, bisogna identificare e scegliere la corretta funzione legame che leghi il valor medio μ_i con il predittore lineare η_i in maniera tale che resti lineare l'effetto delle covariate, ma non necessariamente il legame tra i due fattori.
Esiste, quindi, una funzione $h(\ )$, nota e derivabile, tale che $\eta_i = h(\mu_i) \Longleftrightarrow \mu_i = h^{-1}(\eta_i)$. Alcune funzione legame sono: log, logit, cloglog, identity, probit.
C'è, inoltre, per ogni distribuzione, una funzione $h(\ )$, detta legame canonico, che semplifica i calcoli inferenziali basati sulla log-verosimiglianza. Tale funzione sarà quella che renderà possibile l'eguaglianza $h(\mu) = \theta(\mu)$ ovvero $\eta = h(\mu) = h(b'(\theta)) = \theta$ ossia $h(\ )$ inversa di $b'(\ )$. La funzione legame è sempre adottata come funzione di default in R.

3.3 Un modello per i nostri dati

Tornando ai nostri dati, passiamo ora a stimare i modelli con i quali potremmo in seguito fare le nostre previsioni sul numero totale di voti che Pat Buchanan avrebbe dovuto ottenere nella contea di Palm Beach.

Come abbiamo già discusso nel paragrafo 3.1, il nostro più grande problema nella stima dei modelli di regressione, è la presenza di eteroschedasticità nei dati dovuta alla grande variabilità della popolazione, ovvero alla grande differenza del numero di abitanti tra le varie contee della Florida. L'assenza di eteroschedasticità è uno dei punti fondamentali per la stima di un modello di regressione, e per cercare di superare, o almeno, di arginare questo ostacolo abbiamo moltiplicato, come descritto precedentemente, il predittore lineare η per la variabile N_i (appositamente ricalcolata come descritto nel paragrafo 3.1).

Chiamiamo, quindi, N_i il numero totale di voti registrati e y_i il numero di voti ottenuti da Buchanan nella contea i ed infine, π_i la proporzione totale dei votanti che hanno supportato il candidato riformista. A questo punto, la distribuzione Binomiale ci suggerisce che $\mathrm{var}(y_i) = N_i \pi_i (1 - \pi_i)$. Visto che

sappiamo, osservando anche i risultati della tabella descrittiva vista in precedenza, che il valore complessivo di π_i è circa 0,003%, si può pensare di omettere $(1-\pi_i)$, in virtù del fatto che è ≈ 1. Di conseguenza anche la varianza subisce delle modifiche e diventa $\text{var}(y_i) \approx N_i\pi_i$. Questo ci porta a pensare, tenendo conto del basso valore della probabilità π_i e alla variabile risposta di tipo conteggio, che la distribuzione Binomiale sia approssimabile alla distribuzione di una Poisson di varianza, appunto, $N_i\pi_i$.
Useremo, quindi, queste due distribuzioni per provare a stimare dei modelli lineari generalizzati.

Tra le variabili e le incognite presenti nella stima di un modello con distribuzione appartenente alla famiglia esponenziale, c'è il parametro di dispersione ϕ. Per i due modelli che seguiranno questo fattore è fissato $\phi = 1$; vedremo solo nei capitoli successivi la possibilità di stimarne un valore diverso.

3.3.1 GLM con distribuzione ~ Poisson

La distribuzione di Poisson è particolarmente indicata per le variabili risposta di tipo conteggio, ad esempio il nostro caso, in cui abbiamo il numero di voti ottenuti da un candidato durante delle votazioni, e accetta, come variabile risposta, solo valori interi positivi.

Passiamo ora alla funzione di densità che, partendo da quella generale descritta nei paragrafi precedenti, diventerà:

$$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = \exp\{y\log(\mu) - \mu - \log(y!)\}$$

ovvero, avendo $Y \sim Poisson(\mu)$, avremo che le funzioni $b(\ )$ e $c(\ )$ risulteranno essere: $b(\theta) = \mu = e^{\theta}$ e $c(y;\phi) = -\log(y!)$ con $\theta = \log(\mu)$ parametro naturale e $\phi = 1$ parametro di dispersione. Infine, avremo che $Y \sim EF(e^{\theta},1)$ con $E(Y) = e^{\theta} = \mu$ e $\text{var}(Y) = e^{\theta} = \mu$.
L'ultima cosa da definire è la funzione legame canonica, che sarà anche quella che adotteremo per il nostro modello. Cercheremo, quindi, una funzione $h(\ )$ tale che $h(\mu) = X\beta = \eta$. Avendo che $b(\theta) = b'(\theta) = e^{\theta}$ prenderemo l'inversa di $b'(\ )$ come funzione legame, quindi useremo $h(\ ) = \log(\ )$. Otterremo di

conseguenza $\eta = \theta$ ($\eta = h(\mu) = h(b'(\theta)) = \theta$, presupposto perché la funzione legame diventi canonica) e, riassumendo, potremo scrivere:

$$h(\mu) = \log(\mu) = \eta \quad \text{oppure} \quad \mu = h^{-1}(\eta) = e^{\eta}$$

Tale funzione legame ha l'ulteriore vantaggio di garantire la positività a μ ed è chiamata regressione poissoniana.

Tornando ai nostri dati ed al caso che stiamo studiando dobbiamo affrontare il problema dell'eteroschedasticità. Per risolverlo abbiamo scelto in precedenza di moltiplicare le covariate per il numero totale di voti registrati. Risulterà,dunque,che il modello avrà la forma: $\eta_i = \sum_{j=1} N_i x_{ij} \beta_j$, ma essendo, come detto in precedenza, N_i costante nella sommatoria in j , lo possiamo raccogliere. Di conseguenza $\eta_i = N_i X\beta$, quindi avremo che $\mu_i = N_i \cdot \eta_i$ (con μ_i valore atteso) ma, applicando la funzione legame che abbiamo definito prima e che unisce il valore atteso con il predittore lineare, otterremo $\mu_i = N_i \cdot h^{-1}(\eta_i)$, quindi $\mu_i = N_i \cdot e^{\eta}$ e, di conseguenza, $\log(\mu_i) = \log(N_i \cdot e^{\eta})$, infine $\log(\mu_i) = \log(N_i) + \eta_i$ che sarà la forma finale del nostro modello. Il fattore "scalare" che dunque utilizzeremo sarà per l'appunto $\log(N_i)$.

La stima dei coefficienti e del modello stesso, in R, avviene tramite il comando "glm". In esso si specificano la formula del modello, la distribuzione della famiglia esponenziale che si desidera, la funzione legame ed infine, un possibile fattore per scalare i valori; in particolare nel nostro caso: glm(formula,family=poisson(link="log"),offset=log(totv2)). Come possiamo notare, è lo speciale comando "offset" che inserisce $\log(N_i)$ nel modello. Questo fa si che la stima dei coefficienti non venga influenzata dalla variabile scelta, ma che essa venga addizionata solo in seguito al modello.

Un altro vantaggio a cui provvede il comando "glm" è quello di fornirci la possibilità di determinare quali variabili siano significative, e quali no, all'interno della nostra regressione. Esso, infatti ci stampa in automatico il p-value relativo al test di significatività che ha per ipotesi nulla $H_0 : \hat{\beta}_j = 0$ contro l'ipotesi $H_1 : \hat{\beta}_j \neq 0$. Il valore del test, che sotto H_0 si distribuisce quasi come $N(0,1)$, è calcolato come $t_{oss} = (\hat{\beta}_j - 0) / st.error$.

Attraverso, appunto, questi test (tecnica denominata backward selection) arriviamo a dire che nel modello stimato, con distribuzione Poisson, tutte le variabili, tranne una, sono significative. L'unico fattore che risulta ininfluente ai fini di predire i risultati di Buchanan, è il logaritmo della percentuale di neri (variabile "lblac").

Passiamo, infine, a descrivere il modello ottenuto con i rispettivi coefficienti stimati:

$$\log(buch) = \beta_0 + \beta_1 lpop + \beta_2 whit + \beta_3 lhisp + \beta_4 o65 + \beta_5 hsed + \beta_6 coll + \\ + \beta_7 in\cot + \beta_8 pbush + \beta_9 pbrow + \beta_{10} pnade + \log(totv2)$$

I valori di β_j sono stati calcolati tramite la stima di massima verosimiglianza con equazione di verosimiglianza: $\sum \frac{\partial \ell}{\partial \beta} = \sum \frac{(y-\mu)}{V(\mu)h'(\mu)} x = 0$ che diviene, nel nostro caso, in presenza della funzione legame $\log(\ )$, $\sum \left(y - e^{X\beta}\right)x = 0$.

I valori stimati sono:

β:	0	1	2	3	4	5
Stima:	-3,53461	-0,04109	0,005933	-0,34679	-0,01449	-0,01798
β:	6	7	8	9	10	
Stima:	-0,01057	-0,03332	0,006633	0,189944	0,159991	

Si può notare, a seconda del segno di ogni coefficiente, quale variabile influenzerà positivamente, facendone crescere il valore, e quale negativamente, facendolo diminuire, il risultato di "buch", ovvero il numero di voti ottenuti dal candidato riformista nelle 67 contee.

Andiamo, adesso, a studiare la bontà del modello che abbiamo stimato. Per fare ciò, abbiamo a disposizione due mezzi: la devianza e i residui.

La devianza, in un modello lineare generalizzato, è intesa come il rapporto di verosimiglianza tra il modello corrente ed il modello saturo (inteso come un modello con distribuzione e funzione legame uguale a quello corrente, ma con tutti i parametri inseriti); quindi avendo n parametri (tutti presenti nel modello saturo) e p ($p < n$) parametri nel modello corrente, con le stime di massima verosimiglianza delle funzioni di log-verosimiglianza $l(\tilde{\theta})$ per il modello saturo e $l(\hat{\theta})$ per il nostro modello, la devianza sarà intesa come: $D(y;\hat{\theta}) = 2\phi\{l(\tilde{\theta}) - l(\hat{\theta})\}$.

Se il modello spiegherà bene i dati, e quindi sarà adeguato, il valore della devianza dovrebbe essere piccolo visto che le due $l(\ )$ dovrebbero quasi equivalersi. Inoltre, possiamo dire, che $D \sim \chi^2_{n-p}$.

Nel nostro caso, R, ci stampa,come valore di devianza, 530.92 con 55 gradi di libertà. Confrontata con χ^2_{55} e con un $\alpha = 0.05$ fissato, la devianza risulta troppo grande, e ci porta a rifiutare il modello corrente perché non in grado di spiegare a dovere i nostri dati. Questo risultato, però, ce lo aspettavamo, difatti, con il grande problema di eteroschedasticità legato alla grande differenza di popolazione tra le contee, era prevedibile che anche la devianza ne venisse influenzata.

Passiamo ad analizzare i residui del modello. I residui sono intesi come la differenza tra i valori stimati del modello $\hat{\mu}_i$ e i dati della variabile risposta y_i.

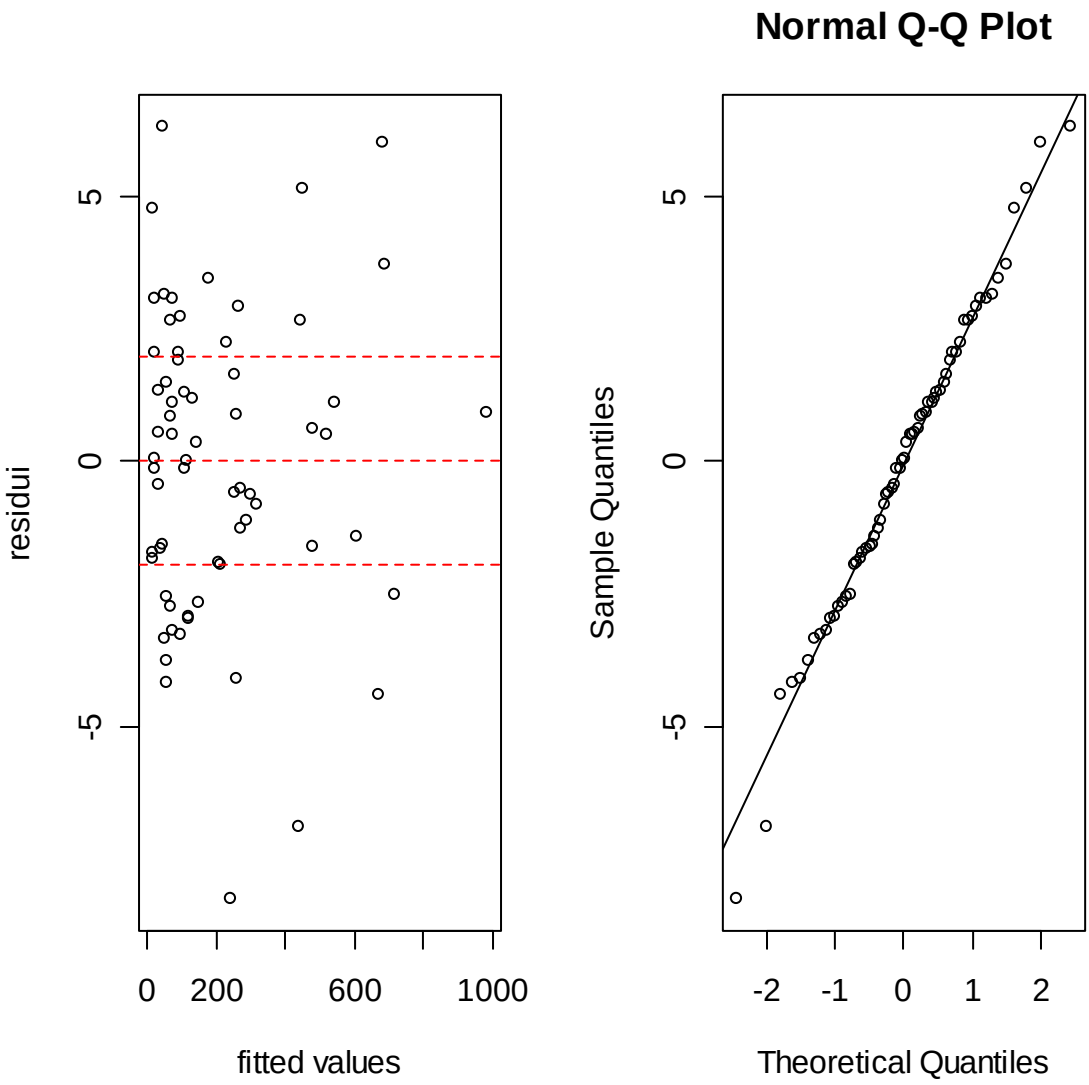
Abbiamo, a nostra disposizione, due tipi di residui (concetti che ampliano i residui descritti per dei semplici modelli lineari): i residui standardizzati di Pearson e i residui legati alla devianza detti, per l'appunto, residui di devianza. Le relative formule sono:

di Pearson: $r_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\phi V(\hat{\mu}_i)}}$ e di devianza: $r_i = \operatorname{sgn}(y_i - \hat{\mu}_i)\sqrt{D_i}$ con $\operatorname{sgn}(\ )$ la funzione segno e D_i la funzione di devianza relativa ad ogni contea i, con $D = \phi \sum D_i$ (in caso di $\phi \neq 1$ i residui vanno riscalati, ma lo vedremo in seguito).

I due tipi di residui sono, comunque, simili $r_P \approx r_D$ e, se il modello è buono, si dovrebbero distribuire come una $N(0,1)$.

Per capire meglio la natura dei residui e il loro livello di bontà, utilizzeremo due grafici: il primo è un diagramma di dispersione tra i residui e i valori adattati al modello, il secondo è un grafico quantile-quantile (“qqplot”) per determinare la normalità dei residui stessi.

Nel nostro caso useremo i residui di devianza, ma gli stessi risultati si otterrebbero con i residui di Pearson.



Al primo grafico sono state aggiunte le linee relative al centro e ai quantili di una normale con $\alpha = 0{,}05$, $\pm z_{1-\alpha/2} = 1{,}96$. Come si può osservare, i residui sembrano essere sufficienti, si approssimano abbastanza bene ad una normale (secondo grafico), anche se c'è una eccessiva variabilità rispetto alla normalità(primo grafico). Questo lo possiamo attribuire, come per la devianza, ad un'eccessiva sovradispersione nella popolazione. Vedremo nel capitolo successivo come superare questo problema stimando un parametro di dispersione più appropriato per spiegare meglio la variabilità. Prima, però, stimiamo un modello con distribuzione Binomiale e $\phi = 1$.

3.3.2 GLM con distribuzione ~ Binomiale

Per questa distribuzione la variabile risposta deve contenere valori compresi tra 0 e 1, $0 \le y_i \le 1$, perché deve esprimere la probabilità di successi sul numero totale di casi, oppure bisogna tener conto sia dei successi che degli insuccessi registrati nello studio del caso. Nel nostro esempio, avendo y_i il numero di voti ottenuti da Buchanan(numero di successi) e N_i numero di voti registrati nelle contee (numero totale di eventi), andremo a creare la variabile proporzione $z_i = y_i / N_i$ che esprime, appunto, la probabilità di successo e i suoi valori sono compresi tra 0 e 1. Questa variabile ci servirà in seguito.

La funzione di densità per $Y \sim Bin(N;\pi)$, con π la probabilità di successo, sarà: $$f(y;\pi) = \binom{N}{y} \pi^y (1-\pi)^{N-y}$$

con $b(\theta) = N\log(1+e^\theta)$ e $c(y;\phi) = \log\binom{N}{y}$; con parametro naturale $\theta = \log\left(\frac{\pi}{1-\pi}\right)$ e parametro di dispersione $\phi = 1$. Infine, a livello generale, $Y \sim EF\left(N\log(1+e^\theta);1\right)$ ed $E(Y) = N\frac{e^\theta}{1+e^\theta} = \pi$ e $\mathrm{var}(Y) = N\pi(1-\pi)$.

Andiamo a definire, ora, la funzione legame appropriata (cercando, come per la Poisson, quella funzione che renda $\eta = \theta$, ovvero la funzione legame canonica). Come abbiamo potuto notare dai risultati sulla funzione di densità mostrati sopra, il valore atteso di Y, μ, comprende anche il fattore N relativo al numero totale di casi studiati. Questa presenza implica due conseguenze: la prima è una trasformazione, la seconda riguarda l'omissione del parametro stesso dalla regressione del modello come fattore scalare.

La trasformazione di cui parlavamo è relativa al passaggio di variabile risposta da y_i a z_i (come spiegato in precedenza). Infatti, passando ad una variabile proporzione, eliminiamo N_i dal rapporto facilitando l'analisi della relazione tra la variabile risposta e la probabilità π_i. In questo modo $E(Z_i) = \mu_i / N_i = \pi_i$.

Questo implica che, fatte le previsioni dal modello stimato, per sapere il numero di voti che avrebbe dovuto ottenere Buchanan, dovremo moltiplicare il risultato (una proporzione compresa tra 0 e 1) per N_i.

La seconda conseguenza la si può spiegare col fatto che, parlando di una variabile che esprime una proporzione, ed essendo nell'ambito delle distribuzioni binomiali, l'influenza di N_i è già presente nel modello e nella relativa regressione. Non dovremo, quindi, preoccuparci, come nel caso della distribuzione di Poisson, di dover moltiplicare il predittore lineare η_i per la numerosità (per arginare il problema della grande variabilità dei dati). Di conseguenza la relazione tra il valore atteso e il predittore sarà $\mu_i^* = h^{-1}(\eta_i)$ (con μ^* valore atteso di Z). Tornando, allora, al calcolo di $h(\ )$ col presupposto che vogliamo $\eta = \theta$ (legame canonico), avremo di conseguenza $\mu^* = \frac{e^{\eta}}{1+e^{\eta}} = h^{-1}(\eta)$ e, invertendo la funzione, $\log\left(\frac{\mu^*}{1-\mu^*}\right) = h(\mu^*) = \eta$.

Tale funzione $h(\ )$, è chiamata regressione logistica o "logit".

Passiamo alla stima del modello e alla valutazione dei vari coefficienti. In R,con il comando "glm", andiamo a creare il modello generalizzato, dovremo però porre, come variabile risposta, una matrice a due colonne, con la prima contenente i successi (la variabile "buch") e nella seconda gli insuccessi ($N_i - y_i$). Risulterà, quindi,

glm(cbind(buch,totv2-buch)~formula,family=binomial(link="logit")). A differenza del modello precedente, non c'è il comando "offset" difatti, come abbiamo detto in precedenza, non scaleremo le covariate per il numero totale di voti.

L'unica variabile che è risultata non significativa (risultato ottenuto tramite la backward selection descritta nel paragrafo 3.3.1) è quella relativa al logaritmo della percentuale di neri (come per il modello precedente).

Il modello sarà:

$$\log it(\mu^*) = \log\left(\frac{\mu^*}{1-\mu^*}\right) = \beta_0 + \beta_1 lpop + \beta_2 whit + \beta_3 lhisp + \beta_4 o65 + \beta_5 hsed +$$

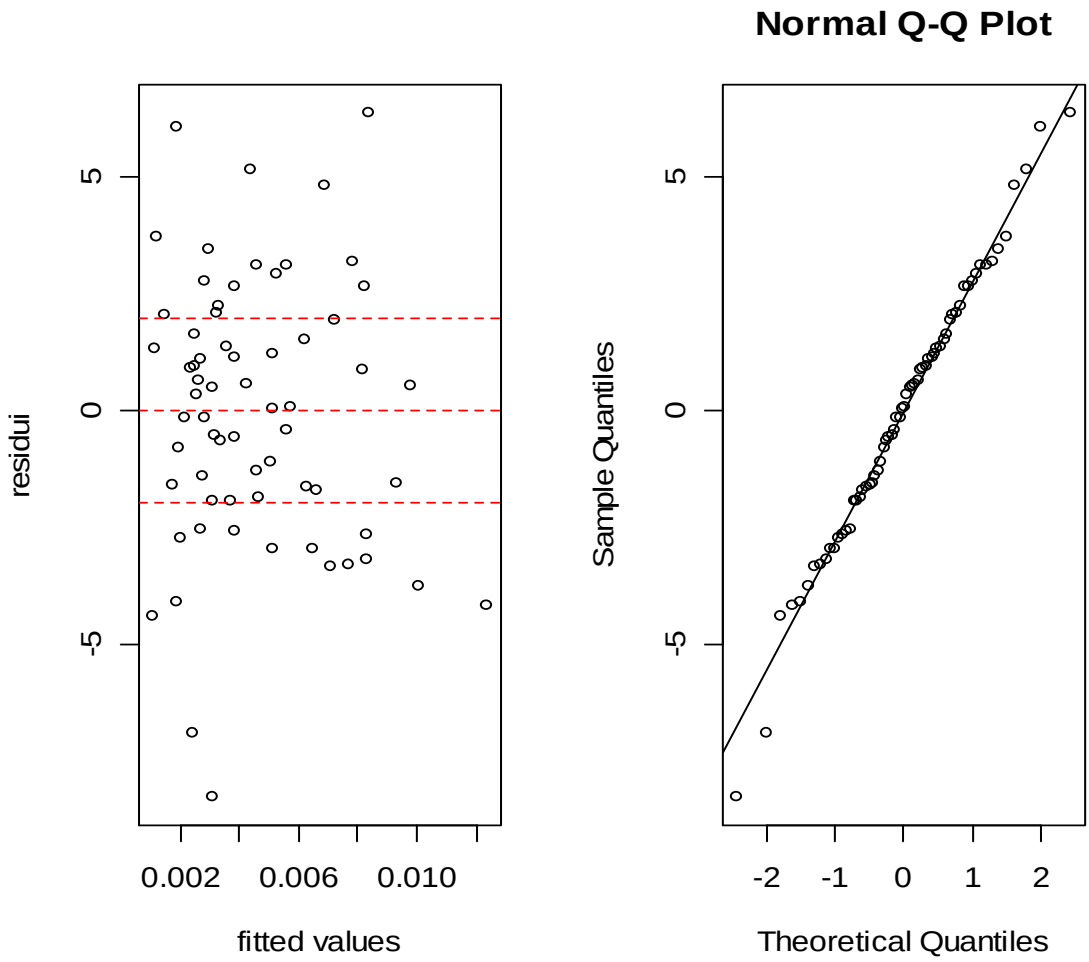
$$+ \beta_6 coll + \beta_7 in\cot + \beta_8 pbush + \beta_9 pbrow + \beta_{10} pnade$$

Con i valori di β_j stimati:

β:	0	1	2	3	4	5
Stima:	-3,51825	-0,0413	0,006003	-0,34814	-0,0146	-0,01814
β:	6	7	8	9	10	
Stima:	-0,01057	-0,03334	0,006619	0,191534	0,160067	

Possiamo notare, dai risultati ottenuti, che i valori dei coefficienti, sono molto simili a quelli registrati nel modello generalizzato con distribuzione di Poisson. Questo è dovuto a fatto che le due distribuzioni, nel nostro caso, sono approssimabili a causa del basso valore di π, 0,003%.

Passiamo allo studio della bontà del modello tramite la devianza ed i residui. Le formule spiegate e descritte nel paragrafo 3.3.1 relativo al modello con distribuzione di Poisson, sono valide anche per questo caso.
La devianza risulta essere $D = 533.13$ e si distribuisce come χ^2 con 55 gradi di libertà. Come con il modello precedente, anche in questo caso tenderemmo a rifiutare il "glm" a causa dell'alto valore registrato, ma, anche in questo caso, la causa è identificabile nell'alta varianza di popolazione tra le varie contee.
Per i residui useremo quelli di devianza e useremo gli stessi grafici presi in considerazione per il precedente modello per analizzarli:



Al primo grafico sono state aggiunte le linee relative al centro e ai quantili di una normale con $\alpha = 0{,}05$, $\pm z_{1-\alpha/2} = 1{,}96$. Anche in questo caso, i residui sembrano essere buoni, si approssimano abbastanza bene ad una normale, come indica il secondo grafico, anche se molti valori escono dalla nuvola centrale descritta nel primo grafico. Come in precedenza e come detto per la devianza, questo lo possiamo attribuire ad un'eccessiva varianza nella popolazione. (Un'analisi condotta sui residui di Pearson avrebbe riportato gli stessi risultati).
Vedremo, ora, come la stima del parametro di dispersione ϕ ci aiuti a diminuire l'eteroschedasticità e a creare modelli più appropriati.

3.4 Modello di Quasi-Verosimiglianza

Può accadere, come nel nostro caso, che l'utilizzo di un modello o che il campionamento casuale di una variabile $y_1 \ldots\ldots y_N$ con una distribuzione come quella Poisson semplice o Binomiale semplice, possa essere inadeguato. Una causa perché questo avvenga è la possibilità che ci sia una maggiore variabilità campionaria rispetto a quella prevista dalla distribuzione. Ovvero che i valori stimati con $\mathrm{var}(Y) = \mu$, per la Poisson, e $\mathrm{var}(Y) = N\pi(1-\pi)$, per la Binomiale, siano inadeguati e sottostimino il reale valore della varianza. Questo fenomeno è chiamato sovradispersione. A livello generale poniamo $\mathrm{var}(Y) = \phi V(\mu)$ con $V(\mu)$ detta funzione di varianza ed è intesa come una funzione del valor medio $V(\mu) = b''(\theta)$ (questa relazione tra la varianza della variabile risposta col valore medio ci permette, e ci ha permesso nei capitoli precedenti, di accettare alcune forme di eteroschedasticità nella stima dei modelli lineari generalizzati). Fino ad ora abbiamo tenuto fisso il parametro di dispersione ϕ e abbiamo utilizzato, sia per il modello distribuito come una Poisson che per quello distribuito come una Binomiale, $\phi = 1$, quindi, con la sovradispersione $\mathrm{var}(Y) > 1 \cdot V(\mu)$, siamo costretti a stimare ϕ per pareggiare questa disequazione.
Per stimare il parametro di dispersione utilizziamo la formula:

$$\hat{\phi} = \frac{1}{n-p} \sum \frac{(y-\hat{\mu})^2}{V(\hat{\mu})}$$

con $n = N$ e $p =$ n°di parametri presenti nel modello corrente.

Quindi, un modello che viene specificato dalle assunzioni $h(\mu) = h(E(Y)) = \eta$, $\text{var}(Y) = \phi \cdot V(\mu)$ e $\text{cov}(Y_k, Y_h) = 0$ per $k \neq h$, è detto modello di quasi-verosimiglianza (con $V(\mu) = 1$ e $h(\mu) = \mu$ il modello è un modello lineare normale).
Un modello di quasi-verosimiglianza è indicato per variabili continue e discrete, in particolare quelle conteggio come nel nostro caso, ed è adatto al superamento dei problemi legati alla sovradispersione che risolve aumentando la varianza $\text{var}(Y)$, oltre a quella fissata dalla distribuzione utilizzata.

Il variare del parametro di dispersione non influenza la stima dei coefficienti lineari β_j del modello, che rimangono uguali a quelli dei modelli con distribuzione della famiglia esponenziale classica (ovviamente la condizione fondamentale perché ciò avvenga, è che, stimando il modello con la quasi-verosimiglianza, si utilizzino solo e tutte le variabili del precedente modello, altrimenti, al loro variare, cambierà anche la stima dei coefficienti β_j).
Cambiano, ovviamente, gli standard error di ogni coefficiente, cambia la devianza e variano anche i residui.

Nel nostro caso, mantenendo tutte le variabili (esclusa “lblac” perché non significativa) otterremo delle stime di ϕ pari a $\hat{\phi} = 9{,}693$, per la distribuzione Poisson, e $\hat{\phi} = 9{,}733$ per la Binomiale. Questo significa, che nei modelli precedenti, avevamo sottostimato la reale varianza di circa 10volte il suo valore originale. Queste stime ci serviranno esclusivamente per gli intervalli di confidenza delle previsione nei capitoli successivi.

Adesso, però, andiamo a stimare dei modelli con la quasi-verosimiglianza, uno con distribuzione Quasi-Poisson e uno Quasi-Binomiale.

3.4.1 Modello ~ Quasi-Poisson

Come abbiamo visto a livello generale nel paragrafo precedente, non sono molte le cose che cambiano dal modello di Poisson stimato nel capitolo 3.3.1. Quindi, dato $Y \sim Poisson(\mu)$, avremo $E(Y) = \mu$ e, soprattutto, $\text{var}(Y) = \phi\mu$. La funzione legame tra μ e η rimane la stessa e, di conseguenza, rimarrà uguale anche il fattore additivo $\log(N_i)$ sommato a fine modello per riscalare i dati. Ciò che cambierà, come già detto, è lo standard error di ogni coefficiente

β_j che verrà moltiplicato per $\sqrt{\hat{\phi}}$; questo implica, necessariamente, che non tutte le variabili utilizzate nel modello precedente rimarranno significative. Il test bilaterale $H_0 : \hat{\beta}_j = 0$ contro $H_1 : \hat{\beta}_j \neq 0$ avrà come valore $t_{oss} = \hat{\beta}_j / \sqrt{\phi V(\hat{\mu})}$ e sarà da confrontare, come in precedenza, con il quantile di una normale $N(0,1)$ con $\alpha = 0.05$ (in R $qnorm(1-\alpha/2) = 1.96$).

Stimiamo, in R, il modello quasi-verosimiglianza per ottenere il valore di ϕ e capire quale variabile influenzerà i voti di Buchanan e quale no.

Il comando informatico resta uguale, basta sostituire il nome della distribuzione poisson, con quasipoisson (il link e offset restano gli stessi come è stato detto prima).

Il parametro di dispersione, per il nuovo modello, viene stimato come $\hat{\phi} = 10.615$.

Le variabili che risultano significative e che andranno a creare il modello sono:

$$\log(buch) = \beta_0 + \beta_1 lhisp + \beta_2 hsed + + \beta_3 in\cot + \beta_4 pbush + \beta_5 pnade + \log(totv2)$$

I valori deu coefficienti sono:

β:	0	1	2	3	4	5
Stima:	-3,889437	-0,353749	-0,018820	-0,049181	0,013325	0,150292

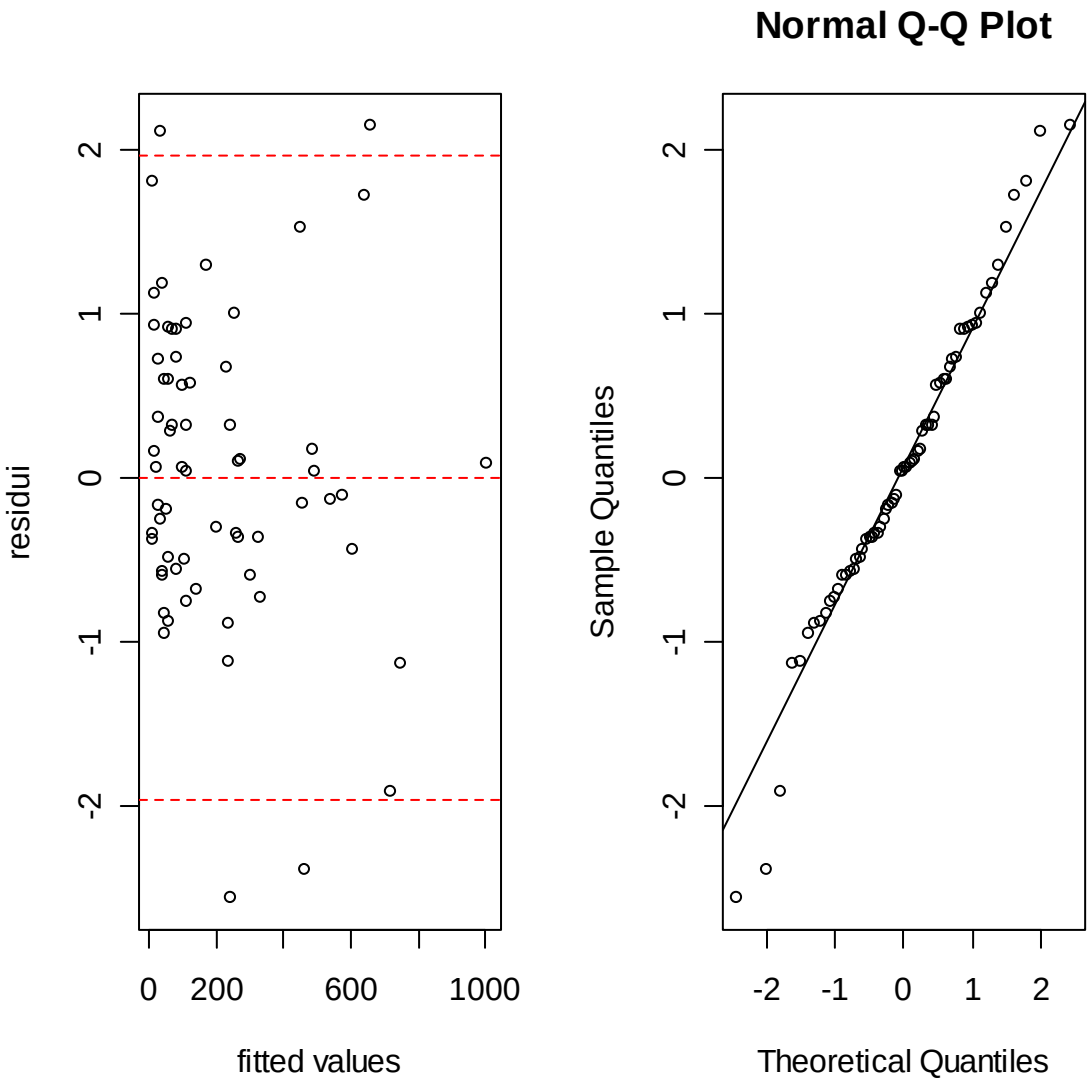
Come ci aspettavamo, molte meno variabili risultano significative.

Passiamo, ora, a valutare la bontà del modello attraverso l'utilizzo della devianza e dei residui, entrambi questi criteri andranno rivalutati rispetto al nuovo valore di $\hat{\phi}$.

La devianza, per la definizione data in precedenza, è $D(y;\hat{\theta}) = \phi \sum D_i$ quindi, per avere il valore scalato, ci basterà dividerla per il parametro di dispersione, $D(y;\hat{\theta})/\hat{\phi}$. Per noi il valore, già diviso per $\hat{\phi}$, è $D = 56{,}94$, che va confrontato con il quantile χ^2_{60} con, appunto, 60 gradi di libertà ($= 79{,}08$ con $\alpha = 0{,}05$).

La devianza risulta minore del valore del χ^2 e, confermato dal valore del p-value del test di 0,59, possiamo dire che il modello sembra buono e sembra spiegare a dovere i dati.

Andiamo ad osservare i residui. Come abbiamo visto in precedenza, anche i residui dipendono dal valore di $\hat{\phi}$; difatti i residui di Pearson, ad esempio, hanno come formula $r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\phi V(\hat{\mu}_i)}}$ solo che prima intendevamo $\phi = 1$, ora, invece, ne conosciamo la stima e quindi la inseriremo nella formula (la stessa cosa vale per i residui di devianza, che , dipendendo dalla devianza stessa, verranno riscalati con essa).
Usiamo i residui di devianza e utilizziamo i grafici, descritti anche per gli altri modelli, per l'analisi:



Possiamo notare che in questi modelli, i residui sono quasi tutti compresi entro i limiti dei quantili di una normale $1-\alpha/2$ con $\alpha = 0{,}05$, e che non hanno valori

estremi, inoltre notiamo che sembrano approssimarsi in maniera abbastanza soddisfacente ad una normale.
In conclusione, possiamo ritenerci abbastanza soddisfatti di questo modello con distribuzione Quasi-Poisson.
Andiamo a stimare un modello Quasi-Binomiale.

3.4.2 Modello ~ Quasi-Binomiale

Per questo modello prendiamo spunto dal modello con distribuzione Binomiale fatto in precedenza, quindi, presa $Y \sim Binomiale(N,\pi)$ avremo $E(Y)=\pi$ e $\text{var}(Y)=\phi N\pi(1-\pi)$. Il cambiamento nella varianza, come lo è stato per il precedente modello di quasi-verosimiglianza, è il principale cambiamento che viene apportato al primo modello stimato con distribuzione Binomiale semplice. La funzione legame rimane la stessa e le motivazioni che ci avevano indotto ad non inserire l'"offset" nel modello pure.

Anche in questo caso il comando R rimane uguale, basterà sostituire "family=binomial" con "family=quasibinomial". La stima del parametro di dispersione la otterremo sempre con questa direttiva R.

Andiamo a vedere quali variabili risultano significative utilizzando la stessa statistica test (riscalata per $\hat{\phi}$) descritta nel modello Quasi-Poisson.
Il parametro stimato risulta essere $\hat{\phi}=10.665$, molto simile al precedente modello, e le variabili significative sono, pure, le stesse di prima (notiamo una certa concordanza tra la distribuzione Quasi-Poisson e Quasi-Binomiale).
Il modello stimato è:

$$\log it(\mu^*)=\log\left(\frac{\mu^*}{1-\mu^*}\right)=\beta_0+\beta_1 lhisp+\beta_2 hsed++\beta_3 in\cot+\beta_4 pbush+\beta_5 pnade$$

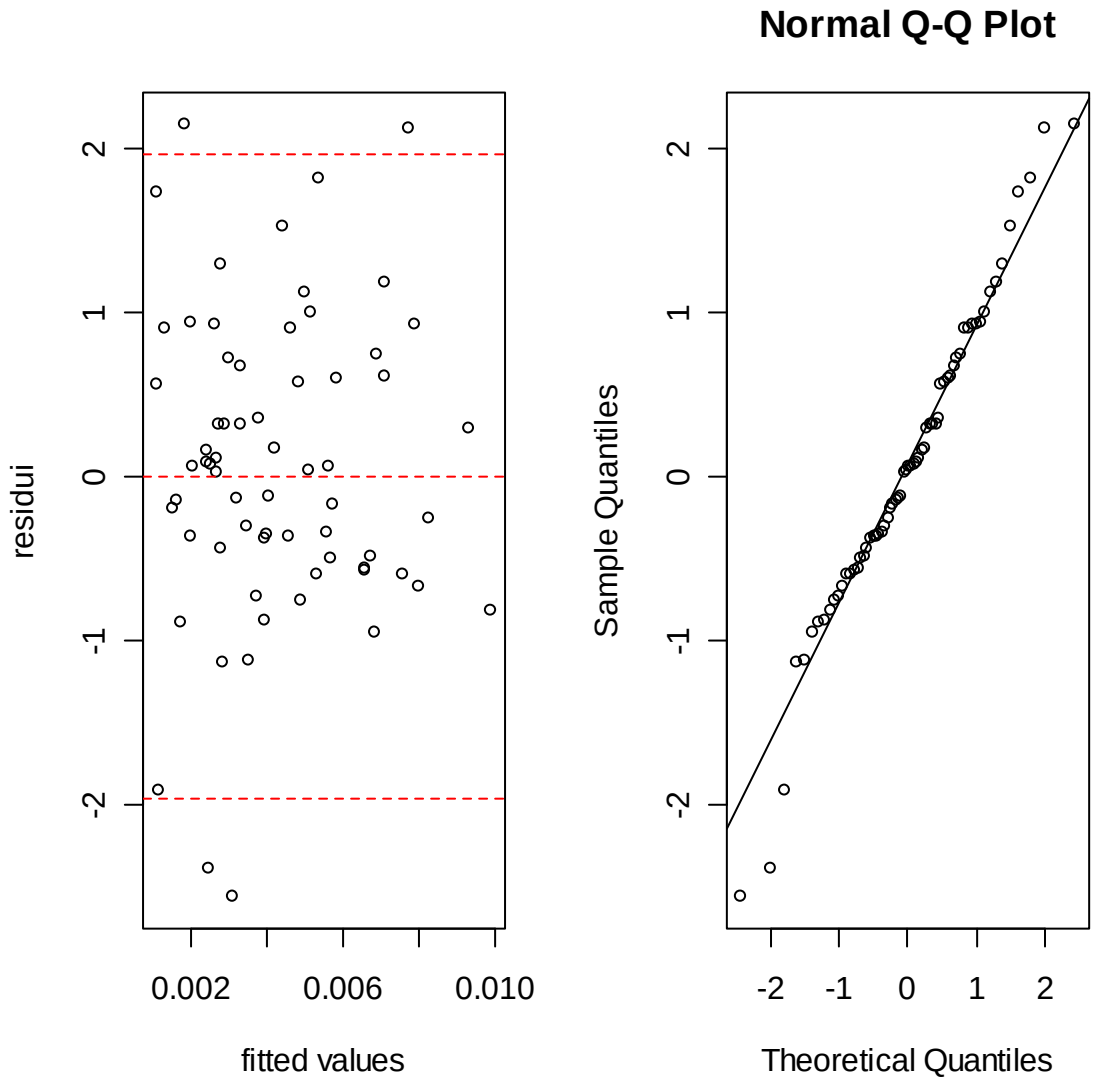
e i valori dei coefficienti sono:

β:	0	1	2	3	4	5
Stima:	-3,876979	-0,354878	-0,018919	-0,049303	0,013361	0,150516

Come per il paragrafo 3.4.1, anche qui useremo la devianza scalata e i residui di devianza scalati come parametro di valutazione del modello. La divisione per $\hat{\phi}$ della devianza e la riparametrizzazione dei residui è uguale a quella fatta con il modello Quasi-Poisson.

La devianza risulta essere $D = 56{,}95$ e, come prima, è maggiore del quantile χ^2_{60} ($D \sim \chi^2$), quindi possiamo dire che i dati sembrano essere ben spiegati dal modello stimato.

Passiamo ad analizzare i residui di devianza attraverso i soliti due grafici già usati per i precedenti modelli:



I commenti che possiamo fare sono simili a quelli già fatti per il modello Quasi-Poisson e possiamo confermare la nostra soddisfazione per il miglioramento dei risultati ottenuti con la quasi-verosimiglianza rispetto a quelli registrati con i modelli normali.

4. PREVISIONI SUL VOTO A PALM BEACH

Giungiamo, infine, al vero obiettivo di questo lavoro, ovvero riuscire, sulla base dei modelli stimati, a fare delle previsioni sul possibile numero di voti che dovrebbe aver ottenuto il candidato riformista Pat Buchanan nella contea di Palm Beach, per poter capire, infine, se i 3407 voti registrati siano, o meno, da considerarsi un risultato erratico, un outlier rispetto al resto dei valori.

Abbiamo creato, per questo scopo, nei precedenti capitoli, quattro modelli con distribuzione Poisson, Binomiale, Quasi-Poisson e Quasi-Binomiale, basati sui risultati registrati nelle 66 contee della Florida, esclusa, dunque, quella di Palm Beach. Andremo, quindi, a fare delle previsioni e a costruire degli intervalli di confidenza per i risultati che otterremo.

Useremo, per calcolare gli intervalli di confidenza, la seguente formula: $\hat{\mu}_i \pm z_{1-\alpha/2} \cdot St.Error$ con un $\alpha = 0.05$, quindi creeremo degli intervalli con livello approssimato al 95% ($z_{0.975} = 1.96$). Ricordiamo, inoltre che, in caso di presenza di una stima del parametro di dispersione $\hat{\phi} \neq 1$, dovremo moltiplicare lo standard error per $\sqrt{\hat{\phi}}$.

Un altro fattore di cui dovremo tener conto è la presenza della funzione legame che unisce il valore atteso μ_i con il predittore lineare η_i, difatti il loro rapporto non è lineare e $\mu_i \neq \eta_i$. Bisognerà, di conseguenza, prestare attenzione alla funzione $h(\ )$ che abbiamo adottato per ogni modello. Noi, infatti, andremo a fare le nostre previsioni tramite il comando R “predict”, che si basa sulla normalità asintotica del predittore lineare, ma che ci stampa la previsione per η_i. Per poter, quindi, ottenere il valore di $\hat{\mu}_i$ (valore atteso della variabile Y_i relativa ai voti ottenuti da Buchanan e scopo del nostro studio) dovremo applicare a $\hat{\eta}_i$ l’inversa della funzione legame $h^{-1}(\ )$. Sempre in R,nel comando “predict”, è presente l’opzione “type=’response’” che consente di fare in automatico questo passaggio.

Andiamo ad osservare la tabella relativa ai risultati ottenuti con le previsioni:

Distribuzione	Valore Stimato	Intervallo Confidenza	Parametro di dispersione $\hat{\phi}$
Poisson	379	(347,410)	1
Binomiale	379	(347,410)	1
Quasi-Poisson	334	(249,417)	10,615
Quasi-Binomiale	334	(249,417)	10,665

Dai primi due risultati notiamo che l'intervallo di confidenza risulta essere molto stretto e limitato per il tipo di previsioni che stiamo facendo e questo lo dobbiamo alla sovradispersione registrata. Proviamo a rifare l'intervallo utilizzando il parametro di dispersione che abbiamo calcolato. Ci basterà moltiplicare lo standard error per $\sqrt{\hat{\phi}}$. I risultati che otterremo saranno di un modello con quasi-verosimiglianza ma con i parametri scelti con distribuzione semplice:

Distribuzione	Valore Stimato	Intervallo Confidenza	Parametro di dispersione $\hat{\phi}$
Poisson (Quasi)	379	(281,476)	9,693
Binomiale(Quasi)	379	(281,476)	9,733

Un'ultima osservazione riguarda i modelli Binomiale e Quasi-Binomiale. Il valore atteso, infatti, lo avevamo inteso come $\hat{\mu}^*$, ovvero relativo alla variabile proporzione $Z = Y / N$. I risultati che abbiamo ottenuto in R, si riferivano, difatti, alla previsione di $\hat{\pi}_i$. Abbiamo registrato, quindi, dei valori espressi in probabilità che poi abbiamo ritrasformato nell'originaria variabile risposta Y_i. Avevamo dunque,nel caso Binomiale, $\hat{\mu}^* = \hat{\pi} = 0{,}000884$, $N = 428879$ e, di conseguenza, abbiamo ottenuto $\hat{\mu} = N\hat{\pi} = 379$ il nostro valore atteso. Lo stesso procedimento lo abbiamo applicato per la Quasi-Binomiale e per il calcolo degli intervalli di confidenza.

5. RIASSUNTO E CONCLUSIONI

In questo lavoro abbiamo cercato di stimare un modello che riuscisse, in maniera soddisfacente, a spiegare i nostri dati. Abbiamo utilizzato la classe dei modelli lineari generalizzati per poter utilizzare distribuzioni della famiglia esponenziale, come Poisson e Binomiale, che ci sembravano più adatte per la nostra variabile risposta, di tipo discreta conteggio. Siamo partiti con le nostre analisi notando subito la presenza di eteroschedasticità nei dati dovuta alla grande differenza nel numero di abitanti per contea. Facendo la stima dei modelli abbiamo dovuto affrontare il problema di sovradispersione nei dati che abbiamo superato con la quasi-verosimiglianza e con la stima del parametro di dispersione ϕ. Alla fine, però, siamo riusciti ad ottenere quattro modelli soddisfacenti ed adeguati.

Analizzando i risultati ottenuti nel capitolo 4, possiamo notare che i due valori principali di predizione sono 379 e 334 e, prendendo gli estremi massimi degli intervalli di confidenza al 95%, otteniamo un intervallo con minimo di 249 voti e un massimo di 476. E' facile notare che il reale valore registrato, 3407 voti, non rientra minimamente in questo range.

Cercando, quindi, di trarre delle conclusioni, possiamo affermare con sicurezza che le elezioni in Florida non si sono svolte con regolarità, o meglio che il cosiddetto "butterfly ballot" ha effettivamente sviato molti voti; perlomeno questo è accaduto nella contea di Palm Beach, a favore del candidato del Reform Party, Pat Buchanan. Egli, infatti, ha ottenuto, nella contea soggetta a questo studio, un surplus di almeno 2800 voti.

In questo lavoro non si è cercato di capire se questi voti "in più" fossero da attribuire ad Al Gore o a Gorge W. Bush (assegnando la vittoria al primo o confermandola al secondo) ma semplicemente si è dimostrato che è successo realmente qualcosa al sistema di votazione, che le macchine punzonatrici non hanno funzionato a dovere o che la scheda "a farfalla" sviava realmente il votante a dare la propria preferenza ad un candidato invece che ad un altro. L'unico punto a favore della teoria che poteva vedere Gore reale vincitore e destinatario di quei voti era la reale maggioranza, circa il 62%, che il partito democratico deteneva in quello stato.

6. APPENDICE

6.1 R

In questo lavoro abbiamo condotto molte analisi per lo studio dei modelli e la descrizione delle variabili. I grafici, le operazioni e i calcoli che abbiamo svolto sono stati fatti quasi[1] tutti tramite il programma statistico R.
Questo software è un insieme di strumenti per la valutazione e lo studio di dati statistici ed è un programma totalmente gratuito reperibile in internet all'indirizzo www.r-project.org. Il fatto di essere un ambiente totalmente Open Source ha reso R, un ottimo mezzo di analisi statistica in continuo aggiornamento.
Andiamo a vedere alcuni comandi che sono stati utilizzati per la stima dei modelli e per particolari grafici o analisi di questo studio[2].
- Il grafico a pag.24 riguardante la relazione tra la percentuale di voti ottenuti da Bush e quella di Gore, con la retta d'andamento:
plot(pbush~pgore) e *abline(lm(pbush~pgore),lty=2,col=2)*
- Stima del modello con distribuzione Poisson:
fitP<-glm(buch~lnpop+whit+lhisp+o65+hsed+coll+incot+pbush+pbrow+pnade, family=poisson(link="log"),offset=log(totv2))
- Stima del modello con distribuzione Binomiale:
fitB<-glm(cbind(buch,totv2-buch)~ lnpop+whit+lhisp+o65+hsed+coll+incot+pbush+pbrow+pnade, family=binomial(link="logit"))
- Stima del modello con distribuzione Quasi-Poisson:
fitQP<-glm(buch~lhisp+hsed+incot+pbush+pnade, family=quasipoisson(link="log"),offset=log(totv2))
- Stima del modello con distribuzione Quasi-Binomiale:
fitQB<-glm(cbind(buch,totv2-buch)~ lhisp+hsed+incot+pbush+pnade,family=quasibinomial(link="logit"))

[1] Gli istogrammi e il grafico a torta delle pagine 19,20,21 sono stati fatti con il programma Microsoft Excel per comodità
[2] Nel capitolo 2 sono elencati i corrispettivi significati delle variabili, le variabili relative ai candidati precedute da una "p", si riferiscono alla percentuale di voto ottenuta dal candidato: pcandidato<-candidato/totv*100 con totv: numero di voti totali.

- Il calcolo e l'analisi grafica dei residui di devianza:

residui<-residuals(fit,type='deviance')*

plot(fitted(fit),residui)* con *abline(h=c(-qnorm(0.975),0,qnorm(0.975)),col=2,lty=2)*

qqnorm(residui) e *qqline(residui)*

- Previsioni sui voti ottenuti da Buchanan (per Poisson e Quasi-P):

predict(fitQB,data.frame(buch),type='response')[50]

- Previsioni sui voti ottenuti da Buchanan (per Binomiale e Quasi-B):

*predict(fitQB,data.frame(cbind(buch,totv2-buch)),type='response')[50]*totv2[50]*

- La creazione della variabile relativa al numero totale di voti "totv":

for(i in 1:67) totv[i]<-

sum(bush[i],gore[i],brow[i],nade[i],harr[i],hage[i],buch[i],mcre[i],phil[i],moor[i])

6.2 I codici delle Contee

Segue la tabelle con la relazione codice - nome di tutte le 67 contee della Florida, il codice è stato assegnato semplicemente ordinando alfabeticamente le contee:

1 Alachua	2 Baker	3 Bay	4 Bradford
5 Brevard	6 Broward	7 Calhoun	8 Charlotte
9 Citrus	10 Clay	11 Collier	12 Columbia
13 Desoto	14 Dixie	15 Duval	16 Escambia
17 Flagler	18 Franklin	19 Gadsden	20 Gilchrist
21 Glades	22 Gulf	23 Hamilton	24 Hardee
25 Hendry	26 Hernando	27 Highlands	28 Hillsborough
29 Holmes	30 Indian River	31 Jackson	32 Jefferson
33 Lafayette	34 Lake	35 Lee	36 Leon
37 Levy	38 Liberty	39 Madison	40 Manatee
41 Marion	42 Martin	43 Miami-Dade	44 Monroe
45 Nassau	46 Okaloosa	47 Okeechobee	48 Orange
49 Osceola	50 Palm Beach	51 Pasco	52 Pinellas
53 Polk	54 Putnam	55 Santa Rosa	56 Sarasota
57 Seminole	58 St. Johns	59 St. Lucie	60 Sumter
61 Suwannee	62 Taylor	63 Union	64 Volusia
65 Wakulla	66 Walton	67 Washington	

7. BIBLIOGRAFIA

- “A statistical assessment of Buchanan’s vote in Palm Beach County”, **Richard L. Smith**, Department of Statistics, University of North Carolina

- “Laboratorio di statistica con R”, **Stefano Iacus** e **Guido Masarotto**

- “Introduzione alla statistica 2: Inferenza, verosimiglianza, modelli”, **Luigi Pace** e **Alessandra Salvan**

- Risultati elettorali dal Florida Department of State, Division of Elections: http://enight.dos.state.fl.us

- Valori demografici dal U.S. Census Bureau: http://www.census.gov/datamap