Izabela Dąbrowska

WMS, rok IV

MU

Numer indeksu: 276351

Maciej Dobrzański

WMS, rok IV

MwI

Numer indeksu: 276733

Model regresji wielokrotnej

Dane wejściowe

Dane do projektu pobraliśmy z biblioteki *faraway* (newhamp). Źródłem tych danych jest publikacja *Herron, M., W. M. Jr, and J. Wand (2008). Voting Technology and the 2008 New Hampshire Primary. Wm. & Mary Bill Rts. J. 17, 351-374*. Zawierają one wyniki w prewyborach prezydenckich w USA w 2008 oraz w 2004 roku w stanie New Hampspire oraz dane demograficzne z tego regionu w podziałem na okręgi wyborcze. Planujemy badać poparcie dla Baracka Obamy w zależności od danych demograficznych oraz wyników prewyborów 4 lata wcześniej (gdzie kandydatami byli John Kerry i Howard Dean).

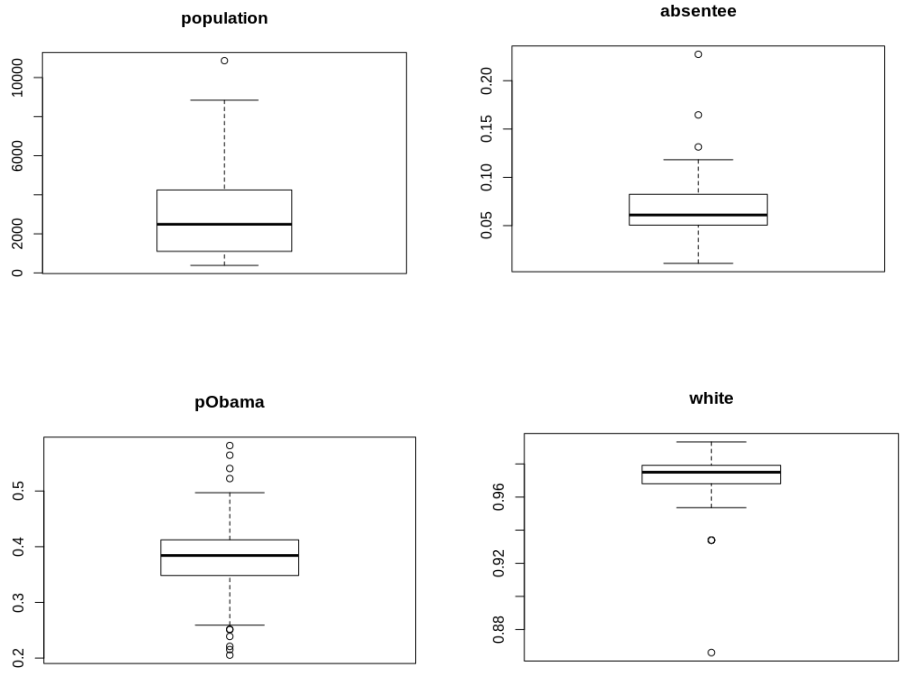
Do analizy wykorzystaliśmy zmienne:

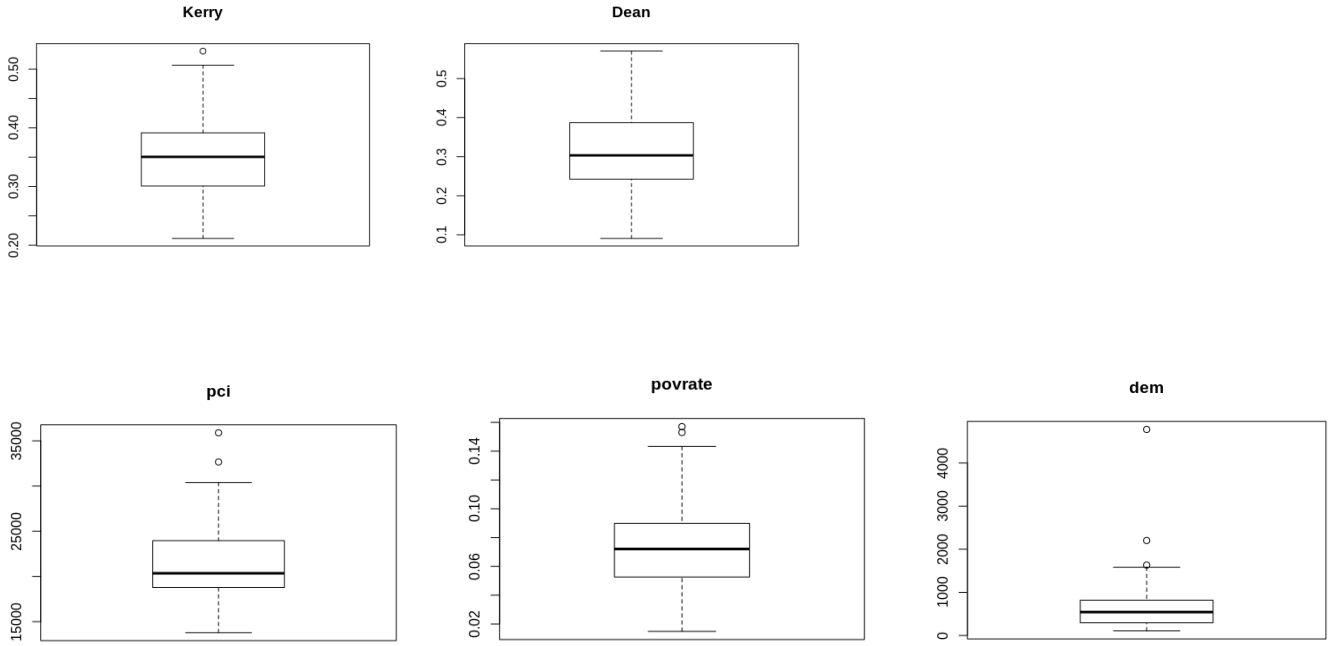
- dem – całkowita liczba głosów na partię demokratyczną w prewyborach
- povrate – stopień ubóstwa w 2000 roku
- pci – dochód per capita
- Dean – procent głosów na Howarda Deana w 2004 roku
- Kerry – procent głosów na Johna Kerry'ego w 2004 roku
- white – procent nielatynoskich białych w 2000 roku
- absentee – procent głosów korespondencyjnych
- population – populacja w 2002 roku
- pObama – procent głosów na Baracka Obamę

Wybór zmiennej zależnej

Naszą zmienną zależną będzie pObama. Będziemy chcieli estymować jego poparcie na podstawie powyższych danych. Postanowiliśmy rozdzielić nasz zbiór danych i wykorzystać pierwsze 99 wierszy do stworzenia modelu, a pozostałe wiersze do przetestowania jego skuteczności.

Wykresy pudełkowe dla poszczególnych zmiennych

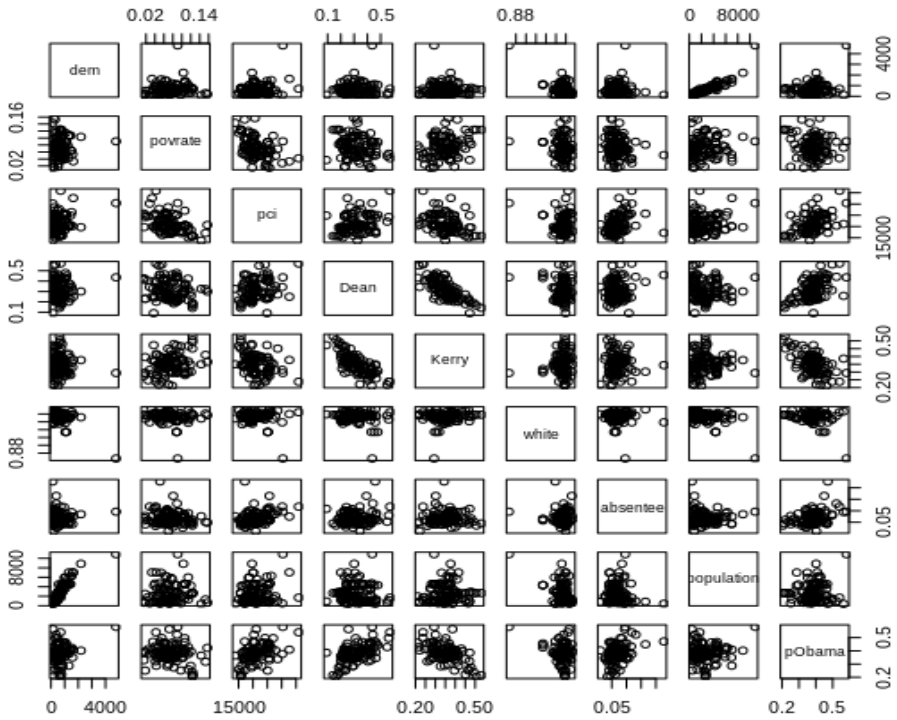




Nie zaobserwowaliśmy dużej liczby obserwacji odstających, niektóre ze zmiennych zostaną poprawione przez transformacje w dalszej części projektu.

Wykresy zależności zmiennych

Na wykresie wstępnie widac, że zmienna zależna powinna być mocno skorelowana ze zmiennymi Kerry oraz Dean, może miec także korelacje ze zmienną pci.



Sprawdźmy macierz korelacji:

	Dem	povrate	Pci	Dean	Kerry	white
dem	1.000000000	0.10089651	0.2950382	0.06013261	-0.042598131	-0.59610827
povrate	0.100896513	1.00000000	-0.4727223	-0.20298227	0.308868281	-0.13071620
pci	0.295038186	-0.47272228	1.0000000	0.30491499	-0.361343018	-0.13930330
Dean	0.060132608	-0.20298227	0.3049150	1.00000000	-0.771744670	-0.15324583
Kerry	-0.042598131	0.30886828	-0.3613430	-0.77174467	1.000000000	0.13637891
white	-0.596108269	-0.13071620	-0.1393033	-0.15324583	0.136378911	1.00000000
absentee	0.001370578	-0.15466811	0.3216854	0.15235002	0.007459074	-0.04893439
population	0.896872729	0.09188685	0.1667668	-0.06590318	0.071717874	-0.42702730
pObama	0.238781821	-0.17401625	0.4906227	0.56080830	-0.572909947	-0.27425831

	absentee	population	pObama
dem	0.001370578	0.89687273	0.23878182
povrate	-0.154668113	0.09188685	-0.17401625
pci	0.321685437	0.16676675	0.49062268
Dean	0.152350018	-0.06590318	0.56080830
Kerry	0.007459074	0.07171787	-0.57290995
white	-0.048934385	-0.42702730	-0.27425831
absentee	0.001370578	0.89687273	0.23878182
population	-0.154668113	0.09188685	-0.17401625
pObama	0.321685437	0.16676675	0.49062268

Macierz wstępnie potwierdza, że poparcie dla Obamy może być opisane przy pomocy zmiennych zależnych, korelacje z nim w większości są dość znaczne (tylko jedna zmienna na poziomie mniejszym niż 0.2).

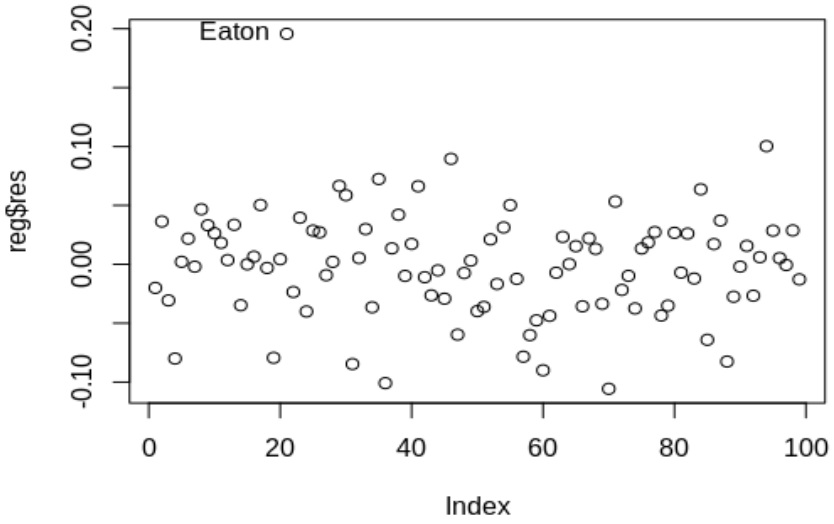
Przeprowadźmy regresję ze względu na wszystkie zmienne:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.114e-01	4.142e-01	1.717	0.089331 .
dem	1.396e-05	2.420e-05	0.577	0.565275
povrate	2.499e-01	1.972e-01	1.268	0.208165
pci	3.744e-06	1.687e-06	2.219	0.028995 *
Dean	9.546e-02	8.291e-02	1.151	0.252677
Kerry	-4.356e-01	1.271e-01	-3.429	0.000917 ***
White	-3.712e-01	4.213e-01	-0.881	0.380653
absentee	7.267e-01	1.864e-01	3.898	0.000186 ***
population	-1.083e-06	6.180e-06	-0.175	0.861278

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04796 on 90 degrees of freedom
Multiple R-squared: 0.5609, Adjusted R-squared: 0.5219
F-statistic: 14.37 on 8 and 90 DF, p-value: 2.648e-13

Sporządźmy wykres reszt i zaznaczmy obserwacje odstające:

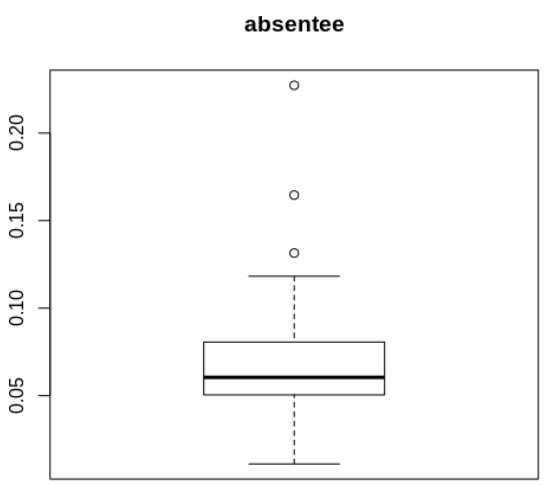


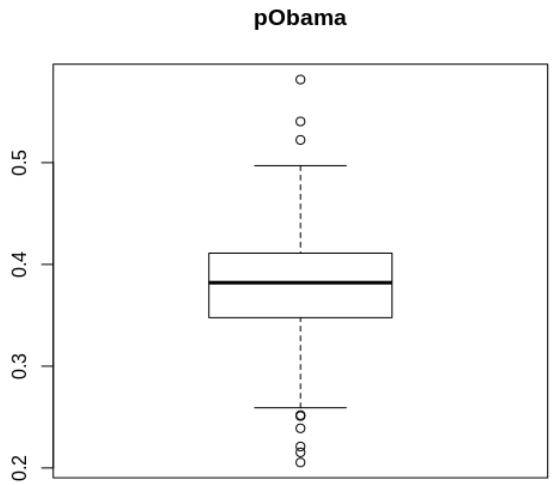
Sprawdźmy normalność reszt:

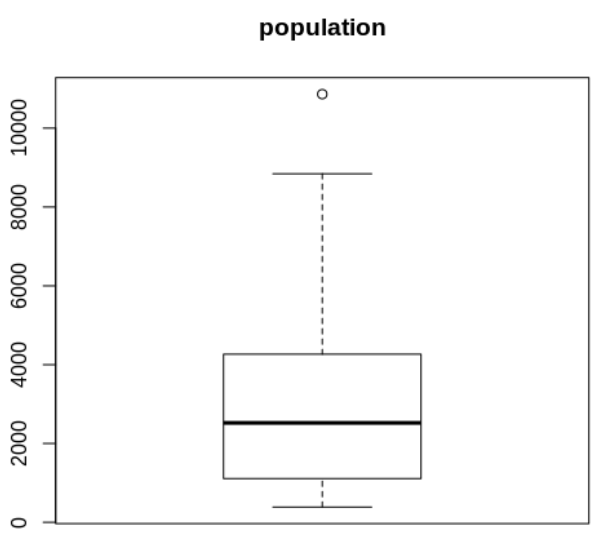
Shapiro-Wilk normality test
data: reg$res
W = 0.96064, p-value = 0.004734

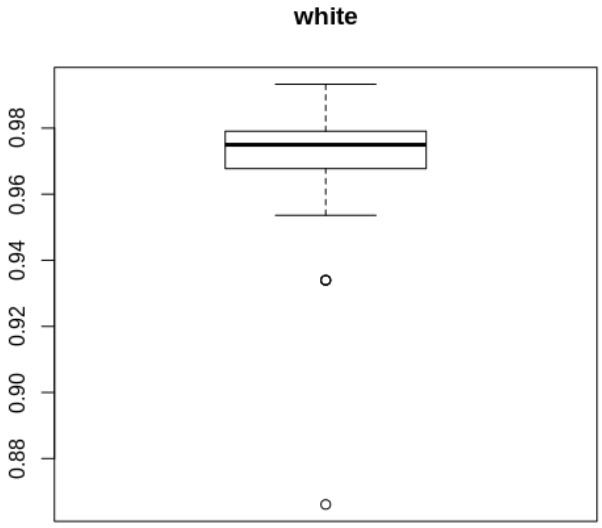
Test wykazał, że należy odrzucić hipotezę o normalności reszt. Analiza wartości odstającej dla reszt, która może mieć na to duży wpływ (okręg Eaton) wykazała, że populacja w tym regionie jest najmniejsza ze wszytkich i stanowi tylko 393 osoby, co może powodować znaczące odkształcenia od modelu. Również ze względu na małą liczbę osób region ten nie powinien być zbyt istotny, dlatego zdecydowaliśmy się odrzucić tę obserwację.

Wykresy pudełkowe po usunięciu Eaton:









Usunięcie nieznacznie poprawiło dane (mniej obserwacji odstających).

Regresja pełna bez Eaton:

	estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.322e-01	3.749e-01	2.220	0.028973 *
dem	6.420e-06	2.191e-05	0.293	0.770189
povrate	2.511e-01	1.780e-01	1.411	0.161869
pci	3.797e-06	1.523e-06	2.492	0.014548 *
Dean	1.157e-01	7.500e-02	1.543	0.126482
Kerry	-4.330e-01	1.147e-01	-3.774	0.000289 ***
white	-5.048e-01	3.815e-01	-1.323	0.189155
absentee	6.667e-01	1.688e-01	3.949	0.000156 ***
population	1.668e-06	5.612e-06	0.297	0.767001

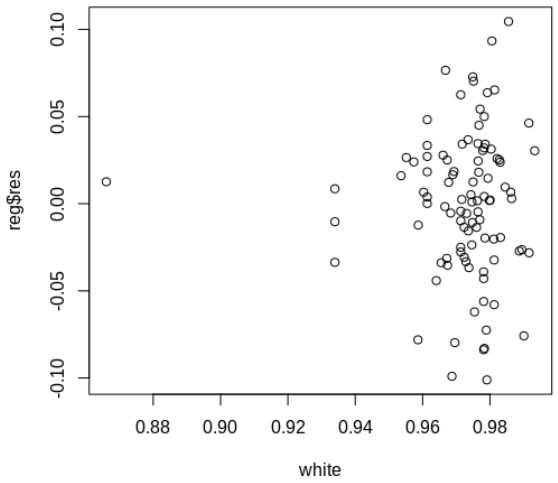
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

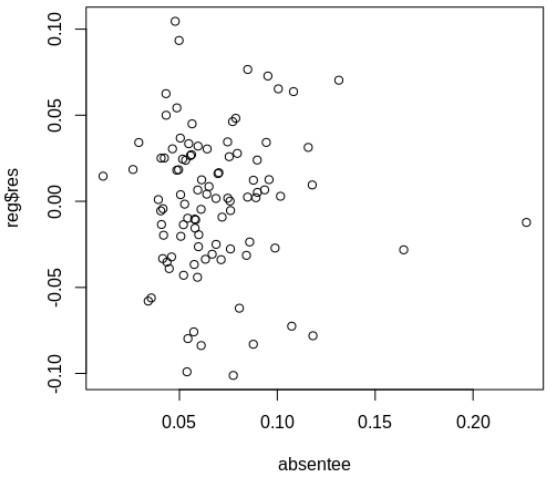
Residual standard error: 0.04331 on 89 degrees of freedom
Multiple R-squared: 0.6187, Adjusted R-squared: 0.5844
F-statistic: 18.05 on 8 and 89 DF, p-value: 9.604e-16

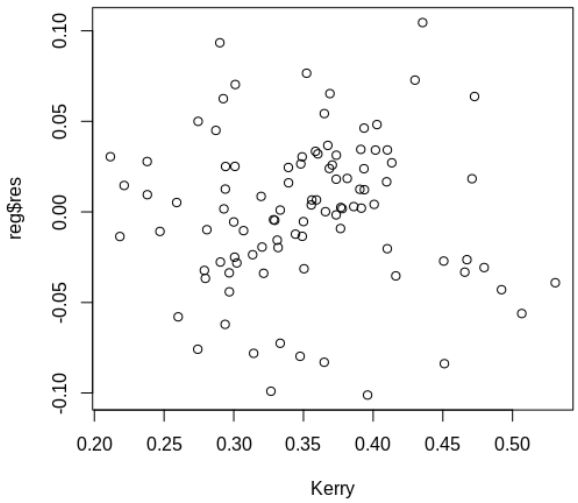
Widać lekki wzrost współczynnika R^2.

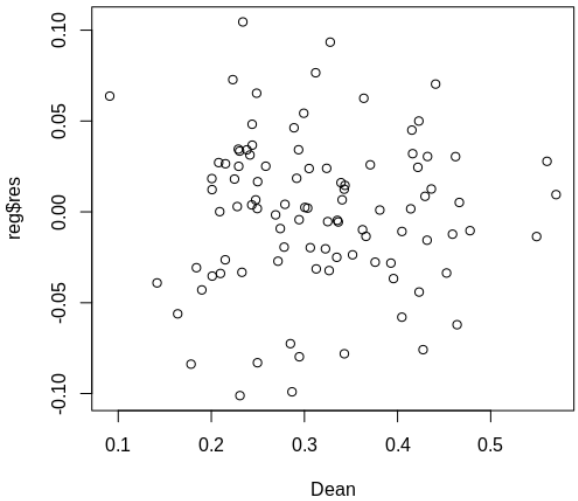
Zbadajmy reszty w naszym modelu

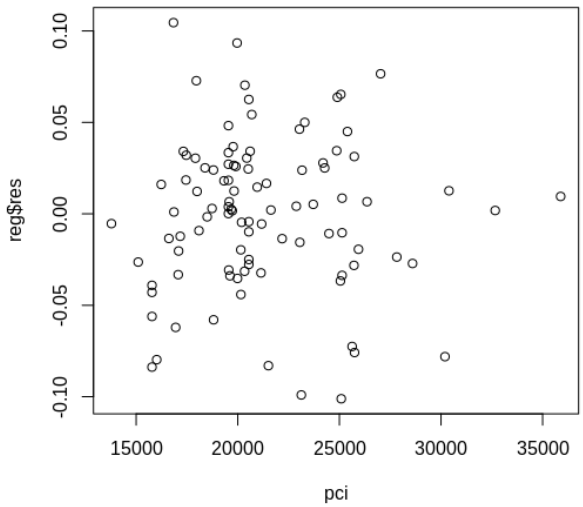
Wykresy zależnośći reszt od zmiennych

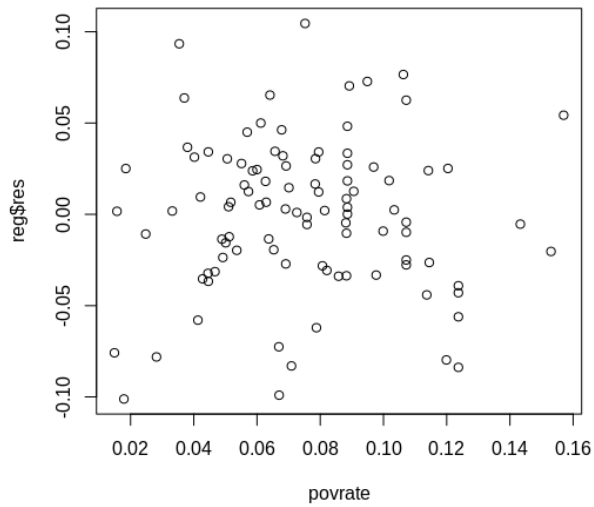


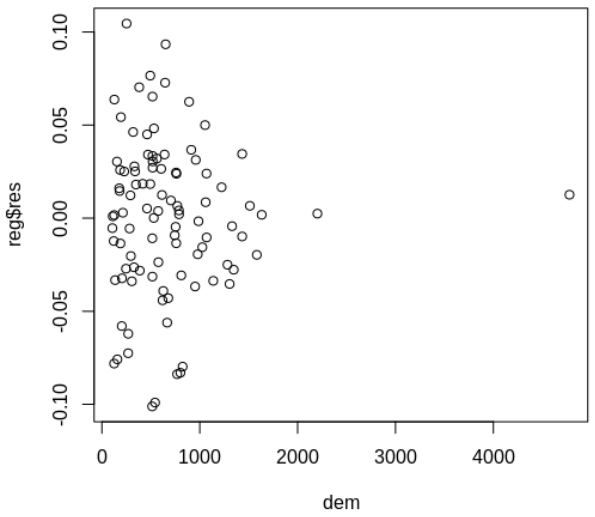


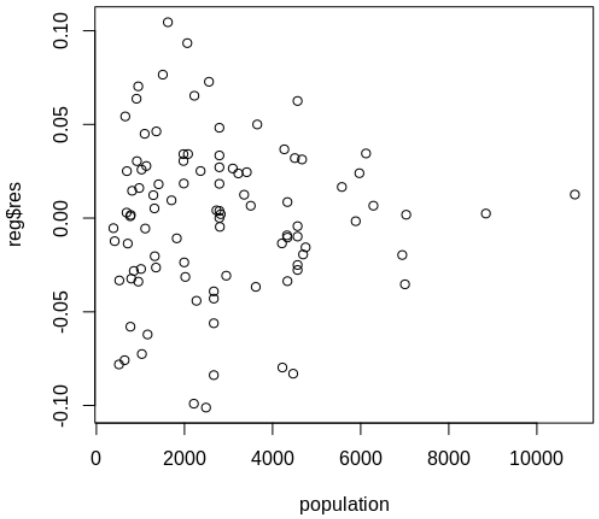


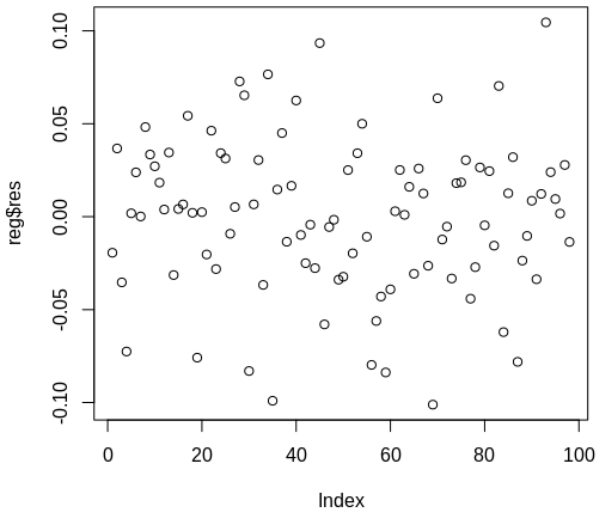






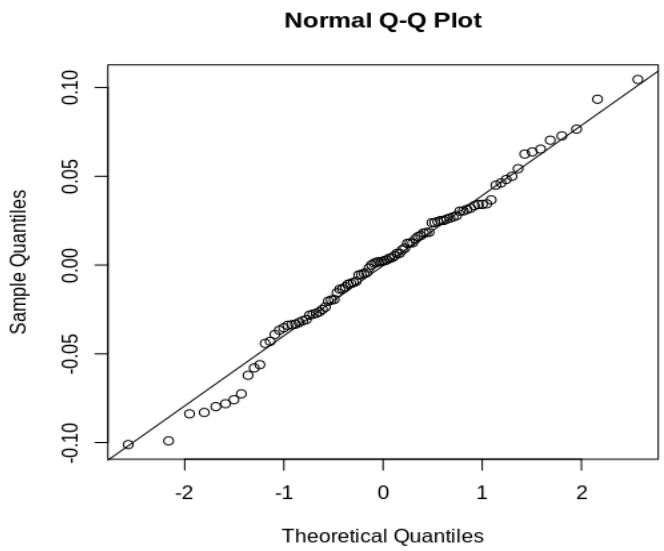






Nie widać zależności reszt od zmiennych zależnych, układają się raczej przypadkowo, symetrycznie wzgledem 0, nie ma też obserwacji odstających.

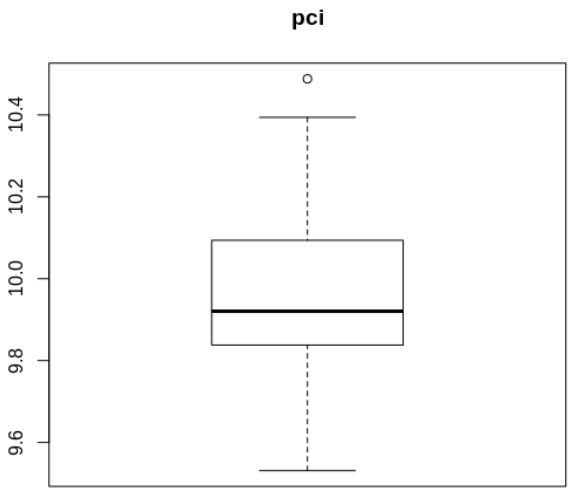
Sprawdzanie normalności reszt

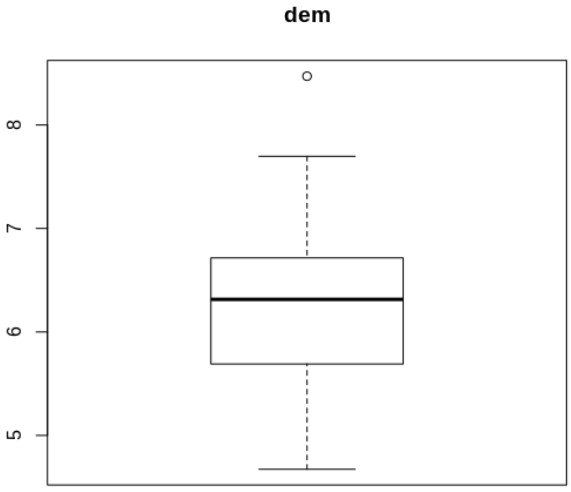


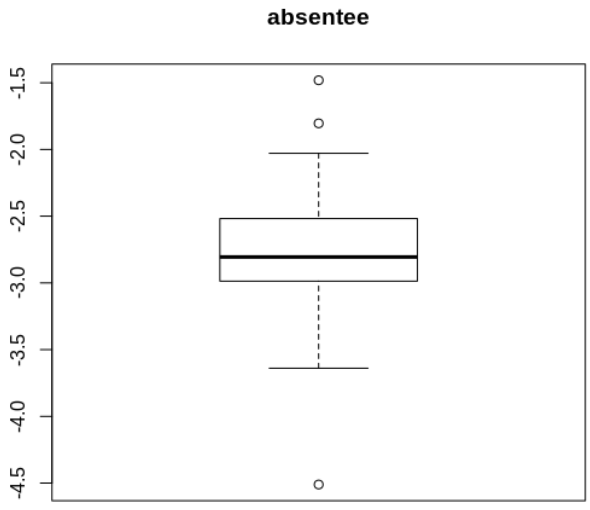
Shapiro-Wilk normality test
data: reg$res
W = 0.98794, p-value = 0.5186

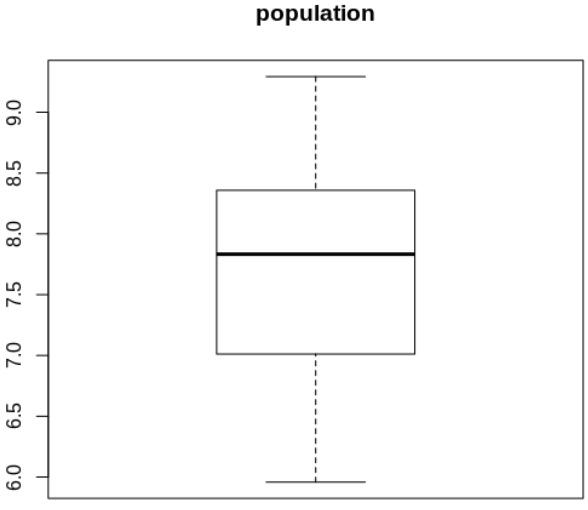
Wykres oraz test Shapiro-Wilka pokazjuą, że teraz nie ma podstaw do odrzucenia hipotezy o normalnosci reszt.

Zajmijmy się teraz potencjalnymi transformacjami zmiennych, poniżej porównania wykresów zmiennych, które zauważalnie poprawiły się pod wpływem transformacji:









Wszystkie transformacje były logarytmiczne, sprawdźmy teraz macierz korelacji:

	log(dem)	Povrate	log(pci)	Dean	Kerry	white
log(dem)	1.00000000	0.1296376	0.2439405	-0.0502431	0.04054441	-0.34043607
povrate	0.12963761	1.0000000	-0.4951880	-0.2039669	0.30982107	-0.13042046
log(pci)	0.24394055	-0.4951880	1.0000000	0.3042418	-0.38598178	-0.13519577
Dean	-0.05024310	-0.2039669	0.3042418	1.0000000	-0.77101941	-0.14827454
Kerry	0.04054441	0.3098211	-0.3859818	-0.7710194	1.00000000	0.13207852
white	-0.34043607	-0.1304205	-0.1351958	-0.1482745	0.13207852	1.00000000
log(absentee)	0.01394683	-0.1367365	0.3407487	0.1185599	0.06050731	-0.08040428
log(population)	0.94581504	0.1220039	0.1203993	-0.1518235	0.14544920	-0.29445511
pObama	0.14717981	-0.1779839	0.5222736	0.6003284	-0.61015350	-0.31260140

	log(absentee)	log(population)	pObama
log(dem)	0.01394683	0.94581504	0.1471798
povrate	-0.13673655	0.12200390	-0.1779839
log(pci)	0.34074873	0.12039929	0.5222736
Dean	0.11855988	-0.15182353	0.6003284
Kerry	0.06050731	0.14544920	-0.6101535
white	-0.08040428	-0.29445511	-0.3126014
log(absentee)	1.00000000	-0.07036157	0.3370630
log(population)	-0.07036157	1.00000000	0.0354780
pObama	0.33706302	0.03547800	1.0000000

Dopasujmy model pełny ze zmiennymi po transformacji

Residuals:

Min	1Q	Median	3Q	Max
-0.106533	-0.027341	0.001458	0.025431	0.103097

Coefficients:

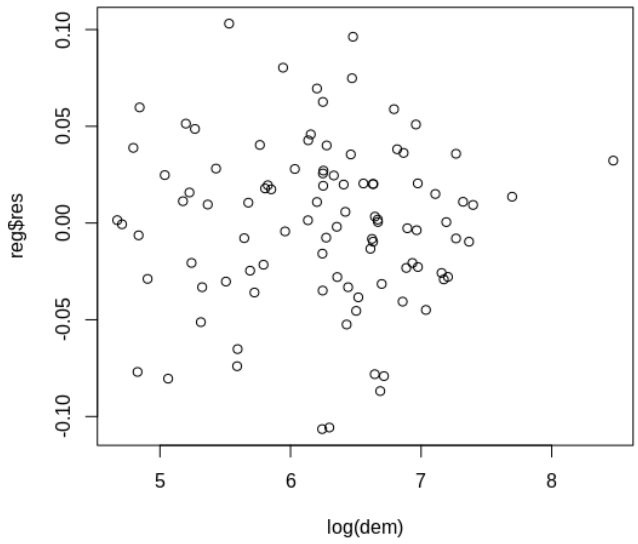
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.16483	0.51496	0.320	0.749659
log(dem)	-0.01016	0.02178	-0.466	0.642170
povrate	0.30258	0.18900	1.601	0.112927
log(pci)	0.10051	0.03612	2.783	0.006585 **
Dean	0.13188	0.07615	1.732	0.086748 .
Kerry	-0.42835	0.12090	-3.543	0.000632 ***
white	-0.63784	0.31936	-1.997	0.048854 *
log(absentee)	0.04477	0.01291	3.468	0.000810 ***
log(population)	0.01415	0.02101	0.674	0.502267

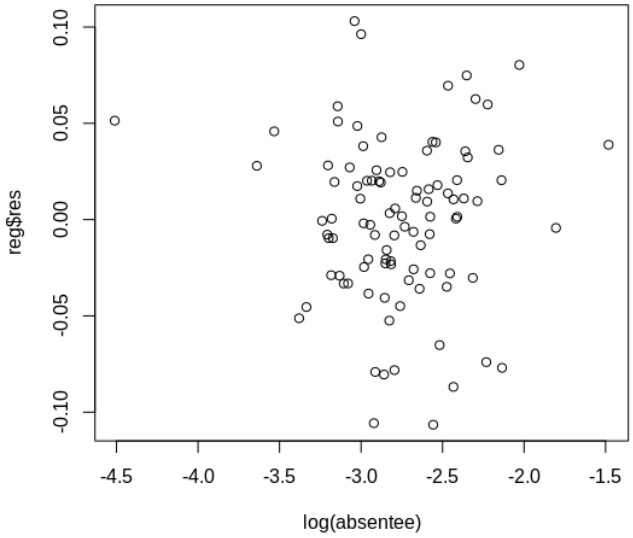
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

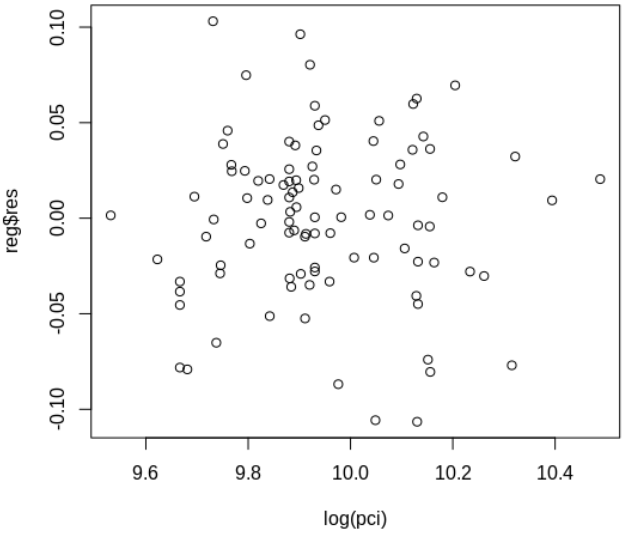
Residual standard error: 0.04383 on 89 degrees of freedom
Multiple R-squared: 0.6095, Adjusted R-squared: 0.5744
F-statistic: 17.37 on 8 and 89 DF, p-value: 2.646e-15

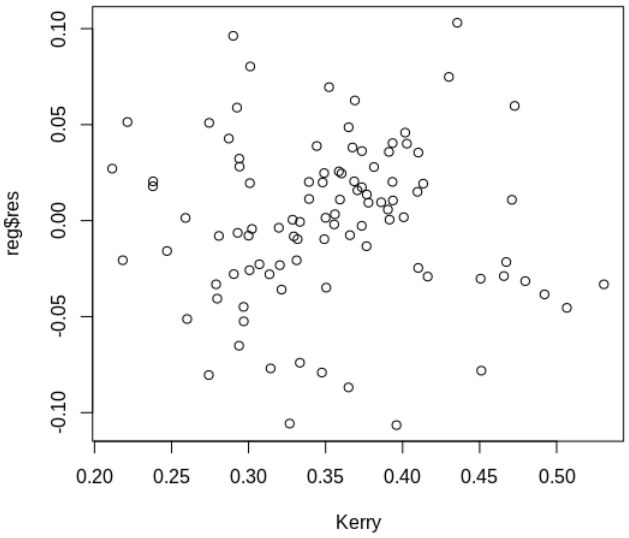
R^2 nieznacznie spadło, za to znacznie wzrosły istotności zmiennych (głównie dzięki ograniczeniu wpływu populacji w danym regionie).

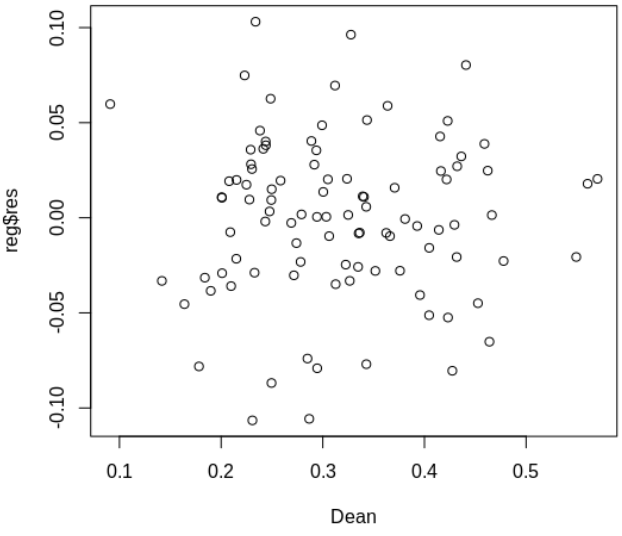
Zbadajmy reszty w tym modelu:

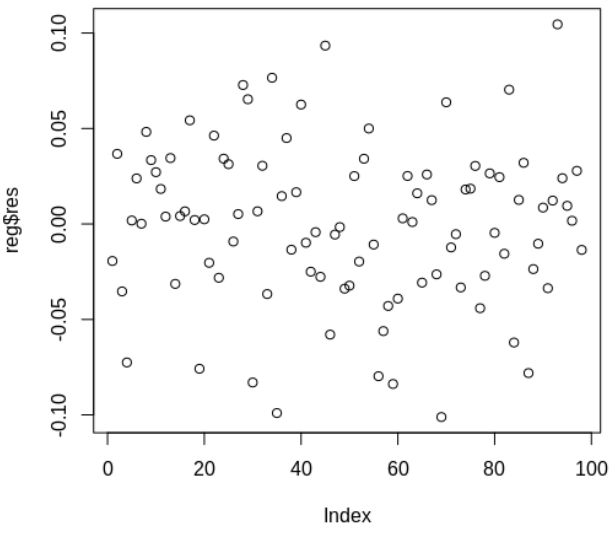


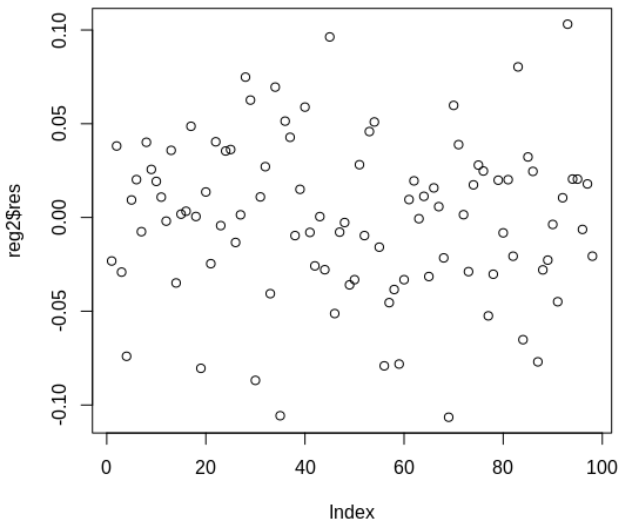


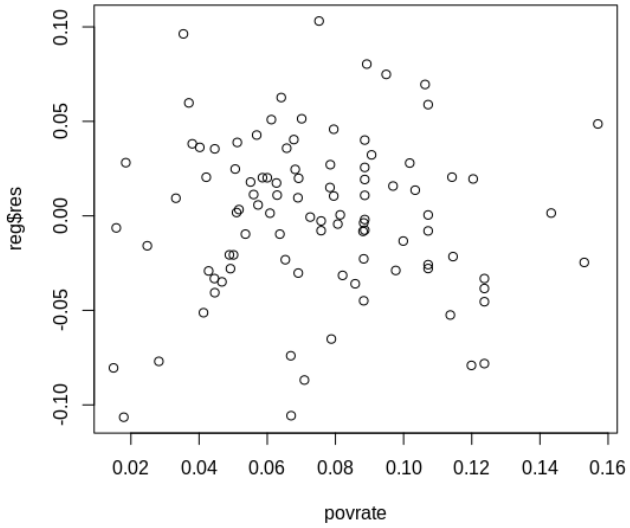




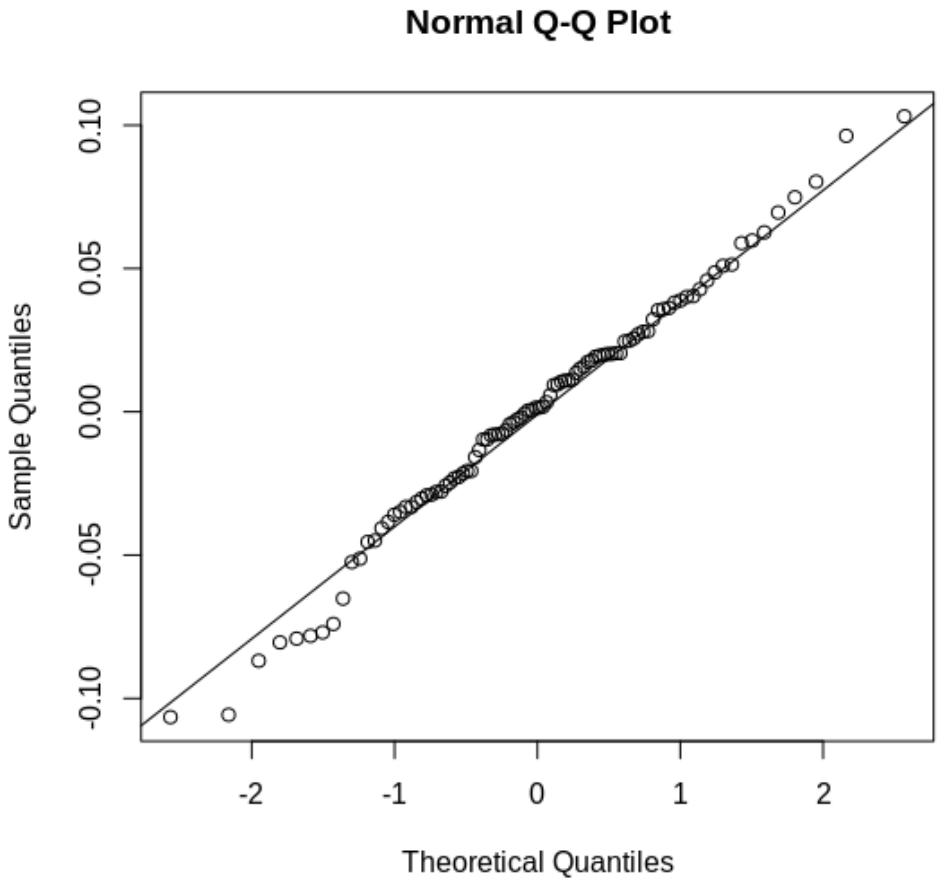








Sprawdźmy normalność reszt



Shapiro-Wilk normality test

data: reg2$res
W = 0.98833, p-value = 0.5476

Widzimy, że w tym modelu również nie ma podstaw do odrzucenia hipotezy o normalności reszt.

Spróbujmy teraz zredukować liczbę zmiennych niezależnych (kryterium Akaike)

tart: AIC=-604.42
pObama ~ `log(dem)` + povrate + `log(pci)` + Dean + Kerry + white +
`log(absentee)` + `log(population)`

	Df	Sum of Sq	RSS	AIC
- `log(dem)`	1	0.0004176	0.17138	-606.19
- `log(population)`	1	0.0008718	0.17183	-605.93
<none>			0.17096	-604.42
- povrate	1	0.0049236	0.17589	-603.64
- Dean	1	0.0057621	0.17673	-603.18
- white	1	0.0076625	0.17863	-602.13
- `log(pci)`	1	0.0148730	0.18584	-598.25
- `log(absentee)`	1	0.0231022	0.19407	-594.00
- Kerry	1	0.0241140	0.19508	-593.49

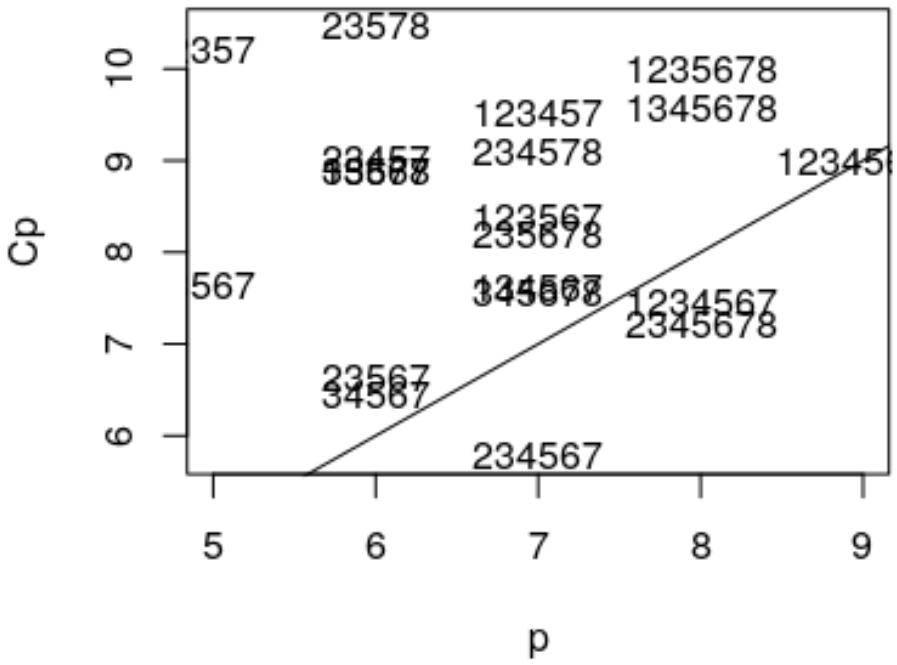
Step: AIC=-606.19
pObama ~ povrate + `log(pci)` + Dean + Kerry + white + `log(absentee)` +
`log(population)`

	Df	Sum of Sq	RSS	AIC
-`log(population)`	1	0.0010604	0.17244	-607.58
<none>			0.17138	-606.19
-povrate	1	0.0045132	0.17589	-605.64
-Dean	1	0.0056971	0.17708	-604.98
-white	1	0.0074531	0.17883	-604.01
-`log(pci)`	1	0.0150835	0.18646	-599.92
-`log(absentee)`	1	0.0226910	0.19407	-596.00
-Kerry	1	0.0236978	0.19508	-595.49

Step: AIC=-607.58
pObama ~ povrate + `log(pci)` + Dean + Kerry + white + `log(absentee)`

	Df	Sum of Sq	RSS	AIC
<none>			0.17244	-607.58
- povrate	1	0.0051422	0.17758	-606.70
- Dean	1	0.0054766	0.17792	-606.52
- white	1	0.0100684	0.18251	-604.02
- `log(pci)`	1	0.0185332	0.19098	-599.58
- `log(absentee)`	1	0.0216571	0.19410	-597.99
- Kerry	1	0.0227600	0.19520	-597.43

Potwierdźmy jeszcze kryterium cp:



Na podstawie powyższych kryteriów stwórzmy nowy model ze zmiennymi : Dean,white,`log(pci)`, `log(absentee), Kerry.

Nowy model:

Residuals:

Min	1Q	Median	3Q	Max
-0.106262	-0.025118	0.000871	0.027491	0.104525

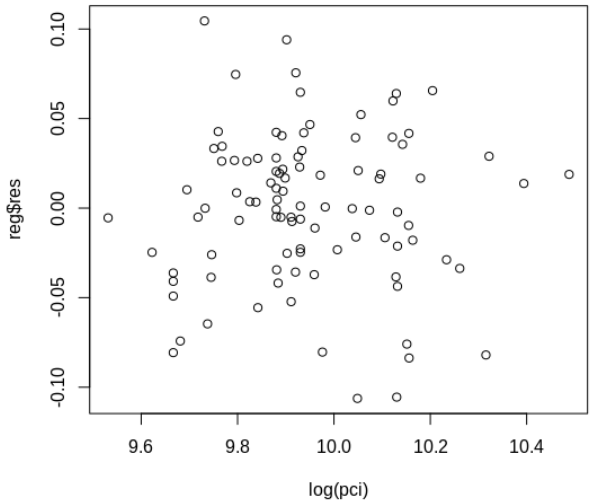
Coefficients:

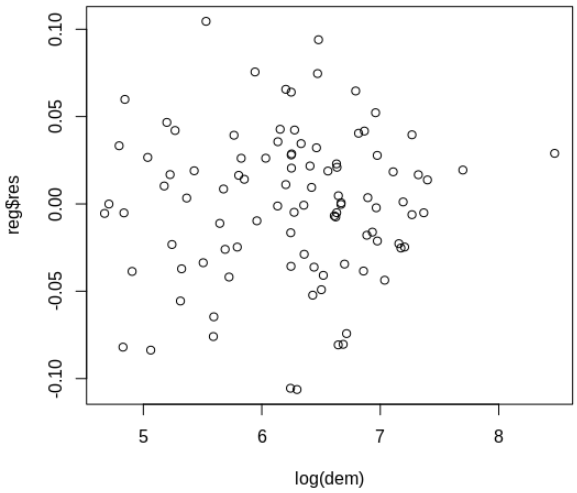
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.25013	0.49606	0.504	0.615310	
povrate	0.28997	0.17603	1.647	0.102945	
`log(pci)`	0.10103	0.03231	3.127	0.002369	**
Dean	0.12839	0.07552	1.700	0.092542	.
Kerry	-0.40802	0.11773	-3.466	0.000809	***
white	-0.69717	0.30245	-2.305	0.023438	*
`log(absentee)`	0.04213	0.01246	3.381	0.001067	**

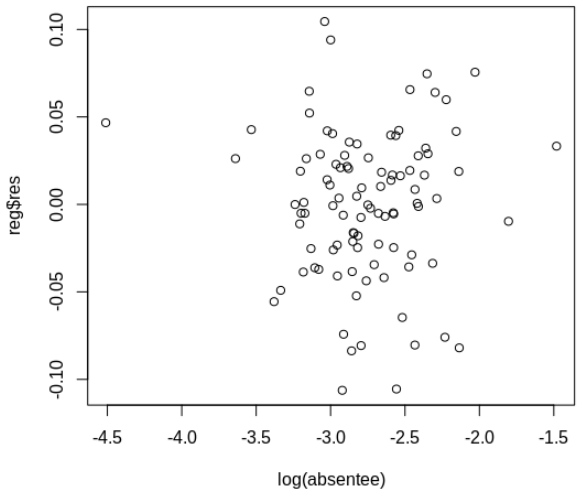
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

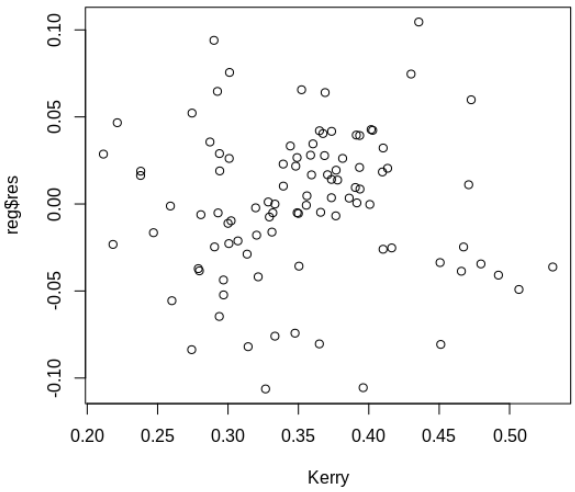
Residual standard error: 0.04353 on 91 degrees of freedom
Multiple R-squared: 0.6062, Adjusted R-squared: 0.5802
F-statistic: 23.34 on 6 and 91 DF, p-value: < 2.2e-16

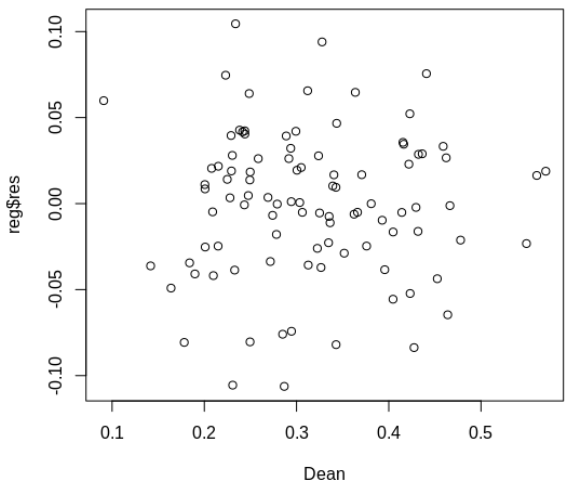
Zbadajmy reszty

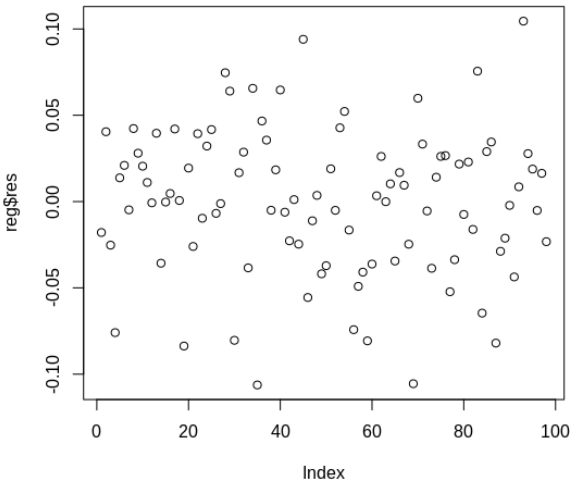


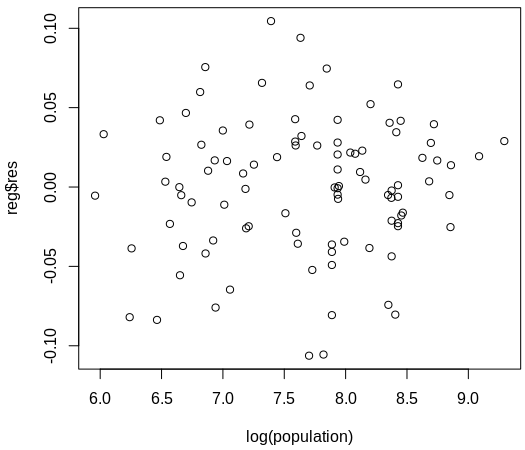


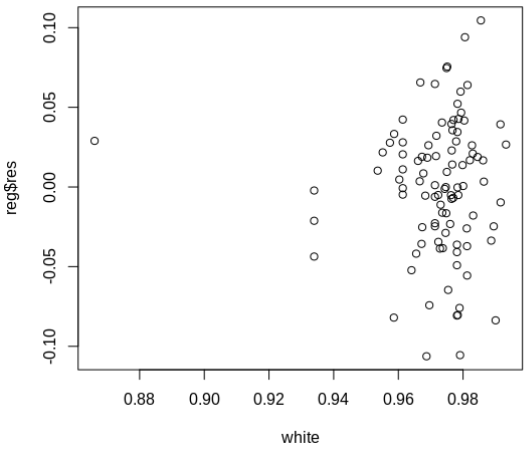


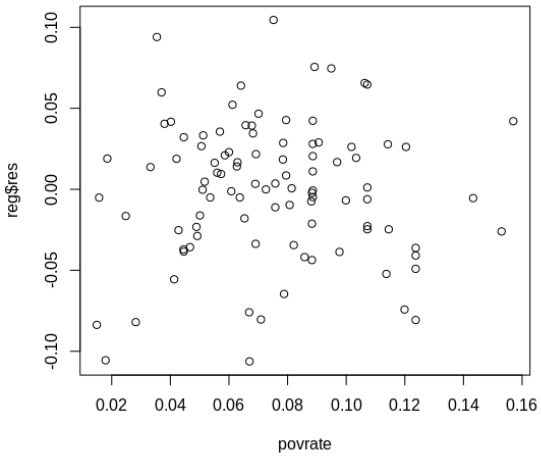




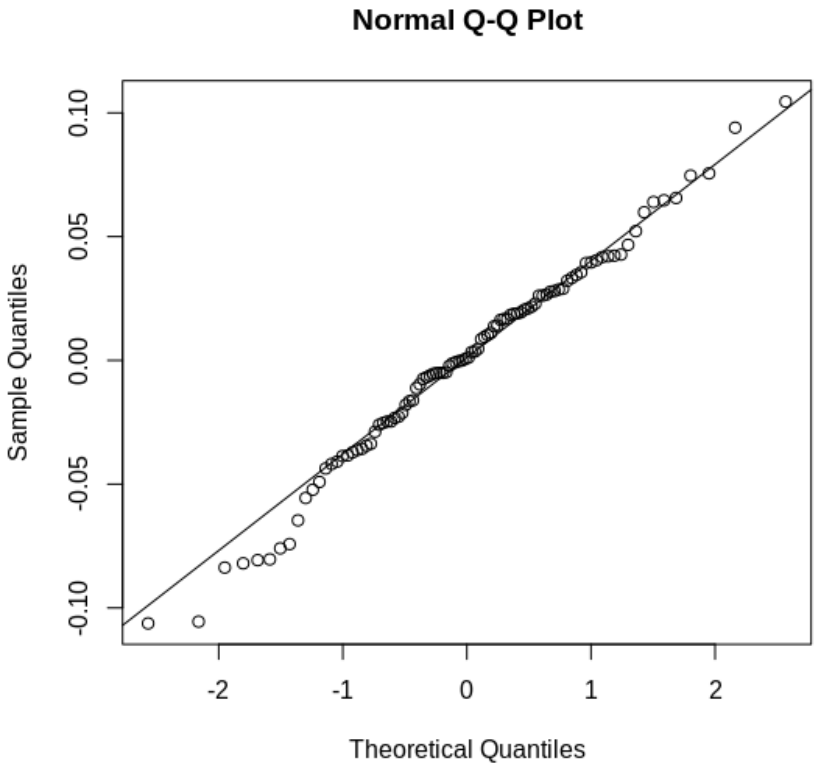








Reszty w nowym modelu wydają się być niezależne od zmiennych zarówno pozostawionych jak i usuniętych.



Shapiro-Wilk normality test
data: reg$res
W = 0.98726, p-value = 0.4705

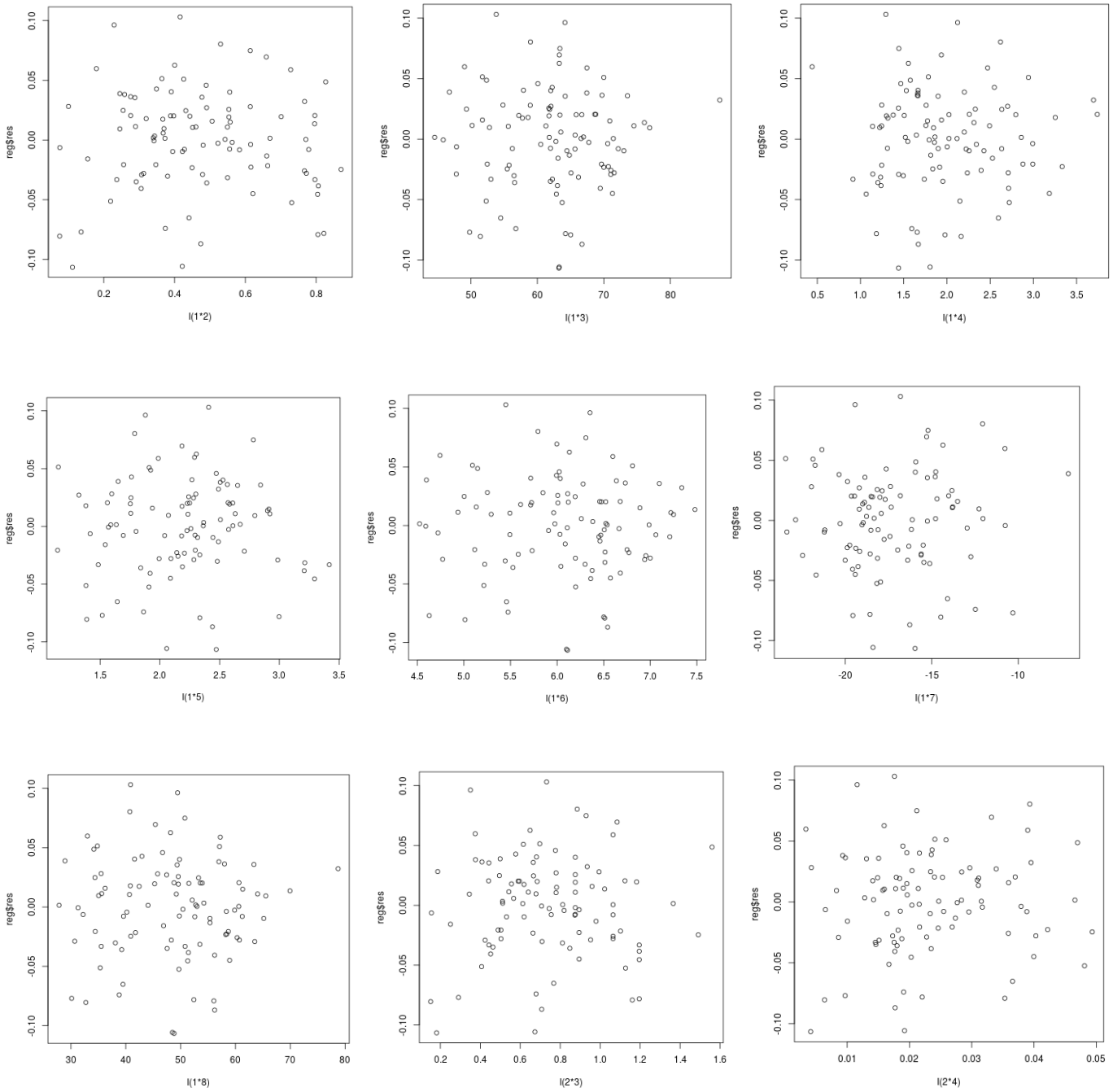
Wykres i test Shapiro-Wilka potwierdzają, że również i tu nie musimy odrzucać normalności reszt.

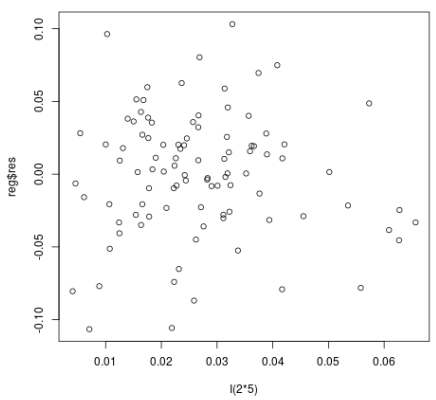
Zbadajmy współzależnośc zmiennych:

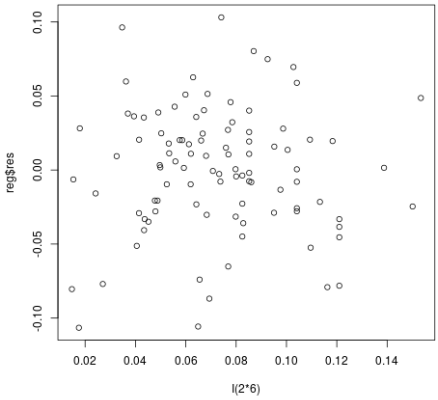
povrate	`log(pci)`	Dean	Kerry	white	`log(absentee)`
1.446590	1.711771	2.691256	3.070138	1.098260	1.292122

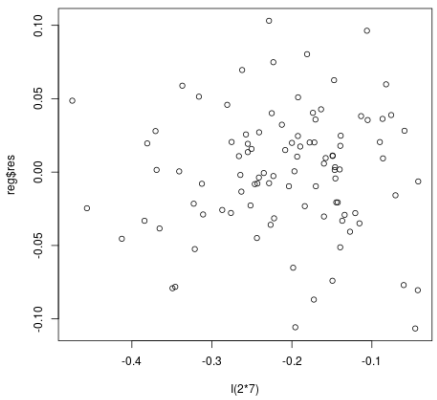
Wszystkie są istotnie mniejsze od 10, więc możemy założyć, że zmienne nie są silnie współzależne.

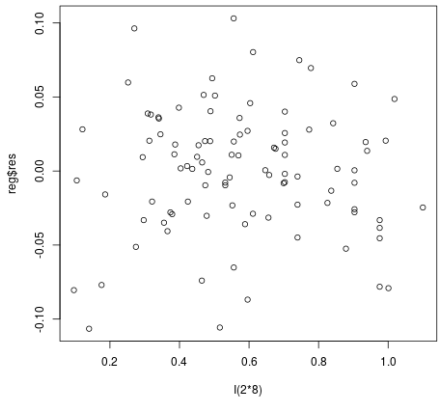
Zbadajmy, czy interakcje pomiędzy zmiennymi mają wpływ na reszty:

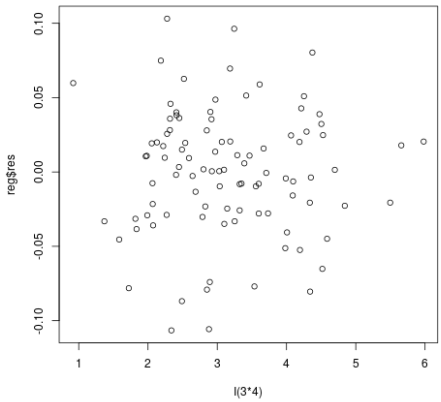


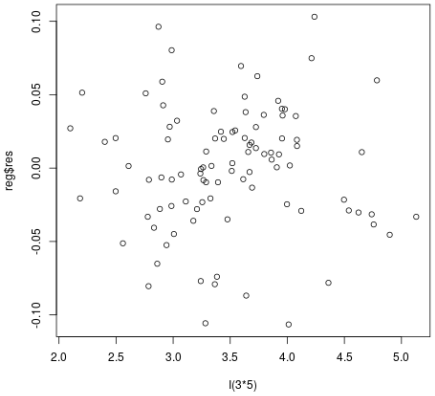

reg$res
I(2*5)

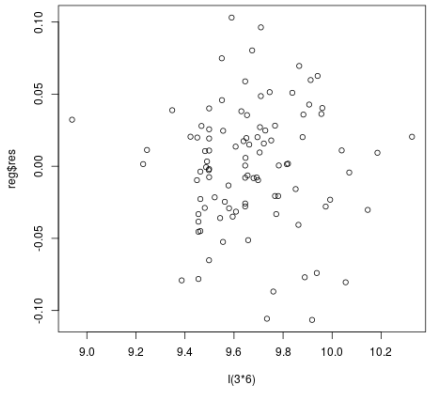

reg$res
I(2*6)

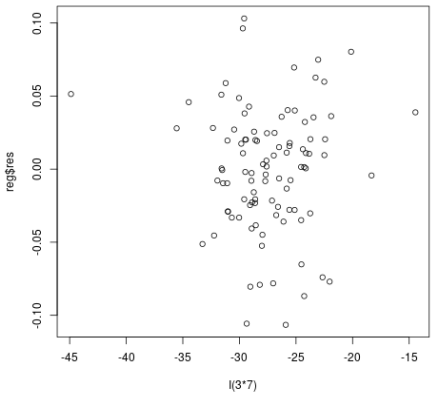

reg$res
I(2*7)

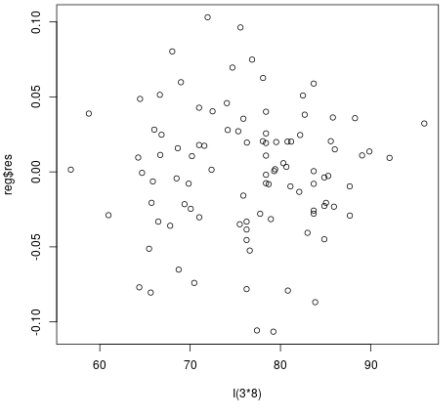

reg$res
I(2*8)

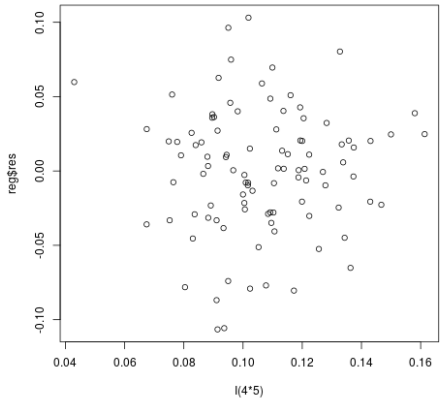

reg$res
I(3*4)

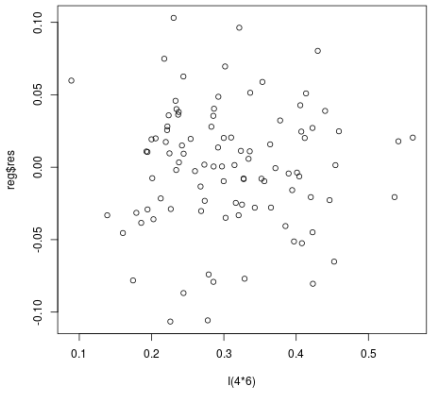

reg$res
I(3*5)

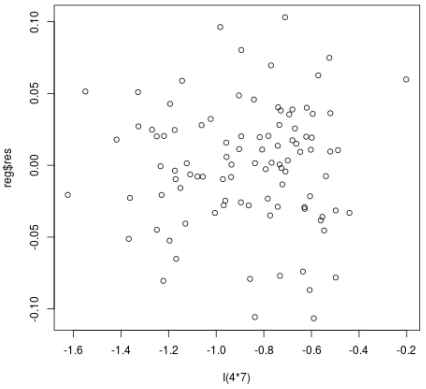

reg$res
I(3*6)

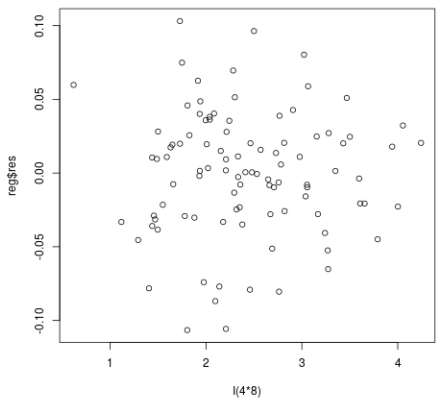

reg$res
I(3*7)

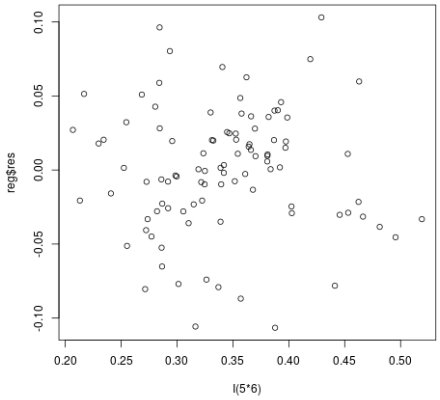

reg$res
I(3*8)

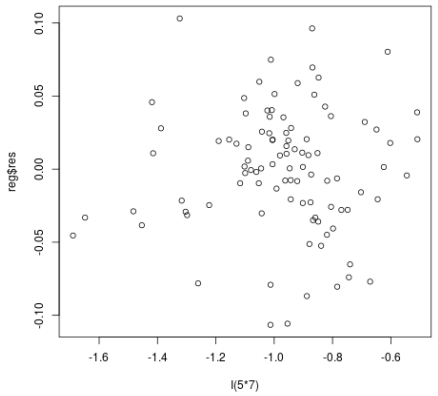

reg$res
I(4*5)

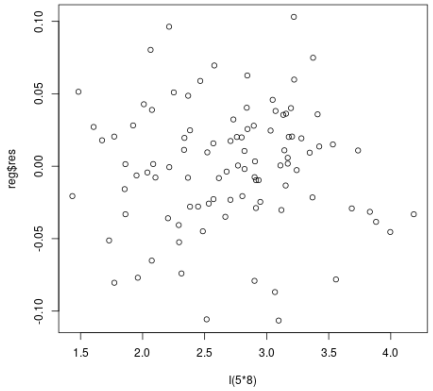

reg$res
I(4*6)

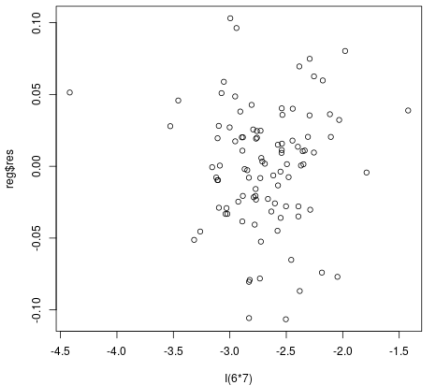

reg$res
I(4*7)

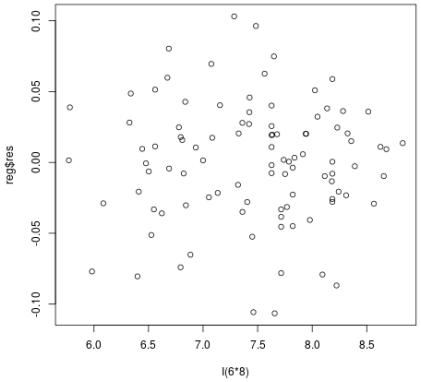


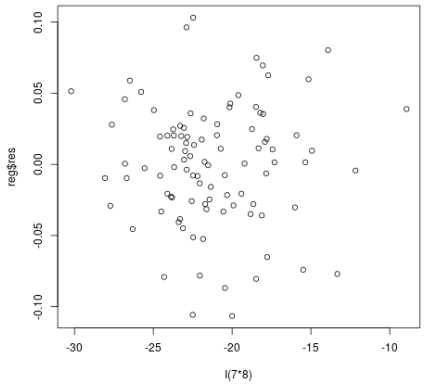












Na powyższych wykresach nie widać wyraźnych zależności reszt od interakcji pomiędzy parami zmiennych, zatem nie będziemy dodawać ich do modelu.

Badanie obserwacji wpływowych

Obserwacje wpływowe

Campton	Rindge	Jaffrey	Fitzwilliam	Ossipee	Bartlett	Conway
0.01702948	0.02152809	0.02192722	0.02495089	0.02675066	0.02696434	0.02775812
LaconiaWard2	Swanzey	Alexandria	Richmond	Wakefield	Chesterfield	LaconiaWard4
0.02869246	0.02997061	0.03029499	0.03030921	0.03164324	0.03170132	0.03215727
LaconiaWard6	Sanbornton	LaconiaWard3	Harrisville	NewHampton	Meredith	Jefferson
0.03239944	0.03418512	0.03631060	0.03635835	0.03642361	0.03811778	0.03910696
Holderness	Lancaster	Freedom	Walpole	KeeneWard4	Tamworth	Gilmanton
0.03933385	0.03990484	0.04132442	0.04285082	0.04309258	0.04402454	0.04486521
Sandwich	Lisbon	KeeneWard1	Littleton	Barnstead	KeeneWard3	Canaan
0.04530211	0.04633197	0.04699767	0.04704635	0.04744170	0.04768310	0.04814874
Tilton	Tuftonboro	Bethlehem	Sullivan	KeeneWard2	KeeneWard5	Enfield
0.04829338	0.04843499	0.04897860	0.04908705	0.05024791	0.05024994	0.05028580
Wolfeboro	Columbia	Madison	Alton	Winchester	Gorham	LaconiaWard1
0.05069594	0.05088319	0.05153870	0.05189458	0.05208332	0.05337065	0.05417785
Westmoreland	Marlborough	CenterHarbor	BerlinWard3	Lincoln	Troy	Hinsdale
0.05424540	0.05458879	0.05485400	0.05760697	0.05767275	0.05841716	0.05879406
Carroll	Bristol	Moultonborough	Haverhill	Belmont	Stark	LaconiaWard5
0.06151287	0.06232338	0.06394462	0.06461003	0.06657010	0.06747375	0.06756024
orthumberland	Milan	Ashland	Brookfield	BerlinWard2	Hebron	Dublin
0.07009159	0.07015823	0.07090506	0.07362919	0.07459031	0.07597059	0.07636368
Marlow	Alstead	Orford	Grafton	Franconia	Piermont	Monroe
0.07702234	0.07897339	0.07903415	0.07911863	0.07952459	0.08312586	0.08531567
Stoddard	Dalton	LebanonWard2	LebanonWard3	Colebrook	BerlinWard1	Effingham
0.09146656	0.09730239	0.09966639	0.09976599	0.10285521	0.10388237	0.10564553
BerlinWard4	Gilford	Bath	Surry	Jackson	LebanonWard1	Shelburne
0.10672735	0.11092695	0.11480111	0.11683886	0.12369293	0.12427787	0.12513098
Pittsburg	Albany	Bridgewater	Lyme	Randolph	Gilsum	Hanover
0.12639601	0.12739782	0.13206576	0.17961716	0.24256217	0.25268742	0.52314622

Decydujemy się pozostawić Hanover, bo jego populacja jest duża, więc ma prawo wpływać na wyniki regresji. Ponadto regresja po usunięciu tego elementu daje nieco gorsze wyniki parametrów R^2 oraz poprawionego R^2 .

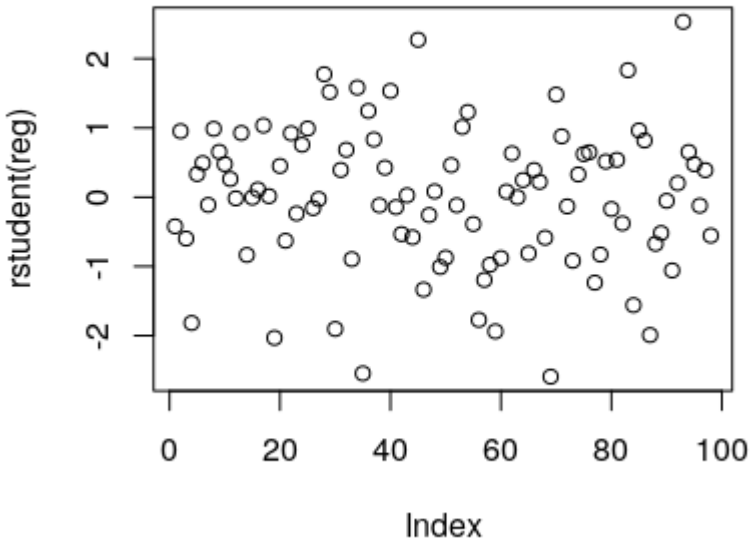
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.04121	0.54166	0.076	0.939531	
povrate	0.28613	0.17615	1.624	0.107795	
`log(pci)`	0.09564	0.03280	2.916	0.004481	**
Dean	0.12408	0.07569	1.639	0.104641	
Kerry	-0.41825	0.11826	-3.537	0.000643	***
white	-0.42160	0.41658	-1.012	0.314225	
`log(absentee)`	0.04237	0.01247	3.397	0.001015	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04355 on 90 degrees of freedom
Multiple R-squared: 0.5698, Adjusted R-squared: 0.5411
F-statistic: 19.86 on 6 and 90 DF, p-value: 1.194e-14

Sprawdźmy studentyzację reszt



Nie ma obserwacji silnie odstających.

Analiza krańcowa

Z powyższych rozważań jako nasz finalny model przyjmiemy:

$$\boldsymbol{pObama = 0.28997 * povrate + 0.10103 * log(pci) + 0.12839 * Dean - 0.40802 * Kerry - 0.69717 * white + 0.04213 * log(absentee)}$$

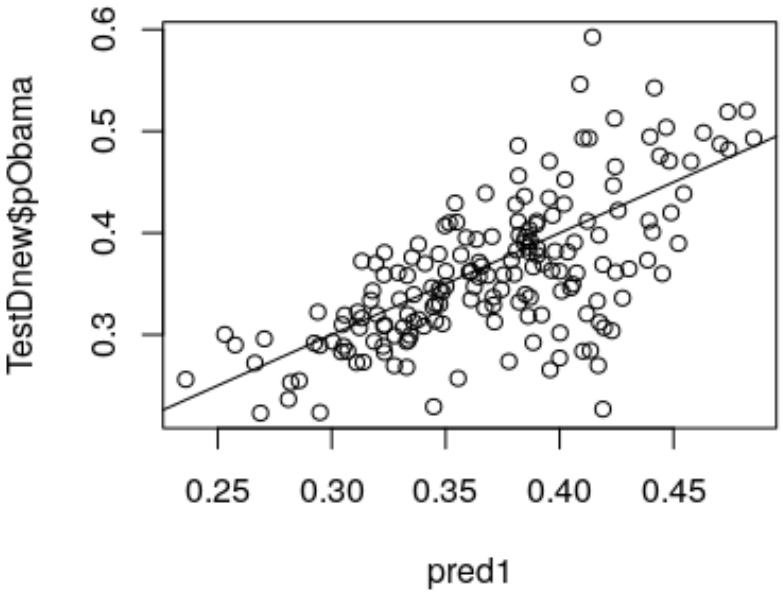
Z analizy naszego modelu wynika, że najbardziej istotną zmienną był wynik Kerry'ego ze współczynnikiem równym -0.408, czyli negatywnie wpływającym na wynik Obamy. Wpływową zmienną był też logarytm ze procenta głosów oddanych korespondencyjnie oraz logarytm z dochodu per capita, obydwa z niedużymi pozytywnymi współczynnikami. Inną wpływową zmienną był procent nielatynoskiej białej ludności, negatywnie skorelowany z wynikiem Obamy. Według podsumowania modelu lekki wpływ mógł mieć też wynik Deana, jednak został on prawdopodobnie przejęty przez wpływ wyniku Kerry'ego.

Możemy go teraz przetestować ze względu na dodatkowe dane, wcześniej przez nas niewykorzystywane:

0.05481714 - średni blad predykcji, uzyskany ze wzoru:

$$\text{Rmse} = \sqrt{\frac{\sum(predicted - test.data)^2}{\text{liczba. wartości}}}$$

Wykres wartości testowych od uzyskanej z modelu predykcji



Wnioski:

Najbardziej wpływowy według podsumowania naszego modelu był wynik Kerry'ego. Jako że zmienna Kerry i zmienna zależna pObama wyrażone są w tej samej skali (jako poparcie), wynika z tego, że znaczna część wyborców Kerry'ego z poprzednich wyborów nie zagłosowała na Obamę, więc wiemy, że zdecydowaną większość wyborców Kerry'ego przejęła Clinton. Jest to związane z podzieałem partii demokratycznej na wewnętrzne frakcje. Podobnie sprawa ma sie z głosami na Deana, jednak zmienna ta została uznana za nieco mniej wpływową, prawdopodobnie ze względu że część informacji którą niosła została już zaaplikowana do modelu przez zmienną Kerry. Pozytywnie na wynik Obamy wpływały zarówno dochód per capita w regionie jak i skala ubóstwa co może sugerować, że otrzymywał on głosy zarówno od najbiedniejszej jak i najbogadszej części społeczeństwa. Negatywny wpływ na jego poparcie miał za tpo procent nie-latynoskiej ludności białej, co może wskazywać, że wyborcy są nieco bardziej skłonni głosować na kandydata z własnej grupy etnicznej. Dodatkowym wynikiem jest dodatni wpływ logarytmu z głosów korespondencyjnych sugerujący, że większa część ludzi którzy nie mogli lub nie chcieli podejśc do urny osobiście wskazała w swoich głosach na Obamę.

Otrzymany model przetestowaliśmy na danych z pozostałych 177 okręgów wyborczych z New Hamspire. Średni błąd szacunku wyniósł około 5 punktów procentowych, zaś linia regresji układała się mniej więcej zgodnie z rozkładem testowych danych (na podstawie wykresu). Można zatem przyjąć że uzyskany model przynajmniej w pewnym stopniu oddaje zależność zmiennej objaśnianej od pozostałych informacji.