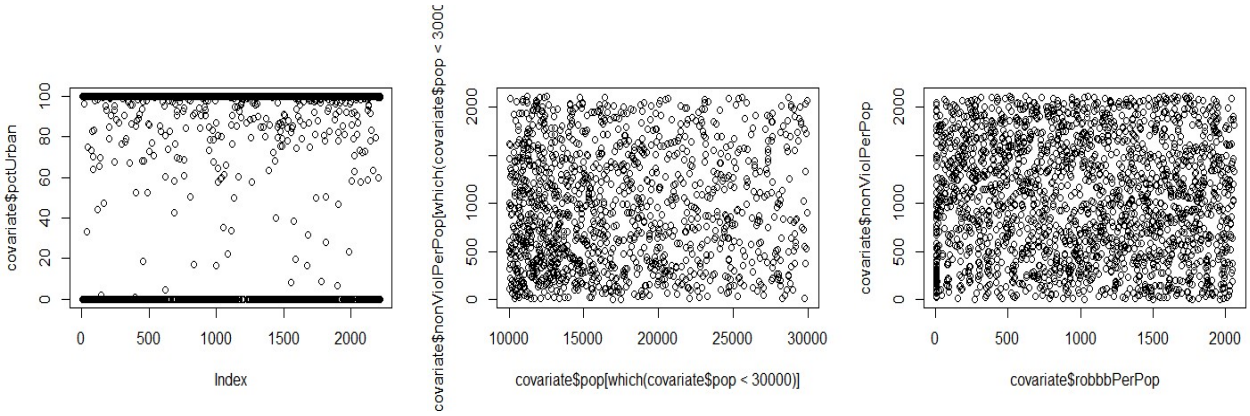
Descrizione del dataset su UCI

```
Communities in the US. Data combines socio-economic data from the '90 Census, law enforce
ment data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data
from the 1995 FBI UCR.

Variables: 125 predictive (1 factor and 124 numeric), 4 non-predictive, 18 potential goal
```

Grafici di alcune covariate senza alcun pattern con il target o distribuzioni degeneri.



Le variabili che hanno almeno un NA ne hanno anche almeno l'80% e son queste.

```
na[na>0.8]
```

```
##      policePerPop policeFieldPerPop   policCallPerPop policCallPerOffic
##         0.8474976         0.8474976         0.8474976         0.8474976
##     policePerPop2       racialMatch     pctPolicWhite     pctPolicBlack
##         0.8474976         0.8474976         0.8474976         0.8474976
##      pctPolicHisp     pctPolicAsian  pctPolicMinority    officDrugUnits
##         0.8474976         0.8474976         0.8474976         0.8474976
##  numDiffDrugsSeiz         policAveOT    policCarsAvail   policOperBudget
##         0.8474976         0.8474976         0.8474976         0.8474976
##    pctPolicPatrol          gangUnit policBudgetPerPop
##         0.8474976         0.8474976         0.8474976
```

Variabili con distribuzioni non attendibili.

```
table(data$persHomeless)['0']
```

```
##    0
## 1637
```

```
table(data$persEmergShelt)['0']
```

```
##    0
## 1234
```

Variabili rimaste dopo la pulizia del dataset.

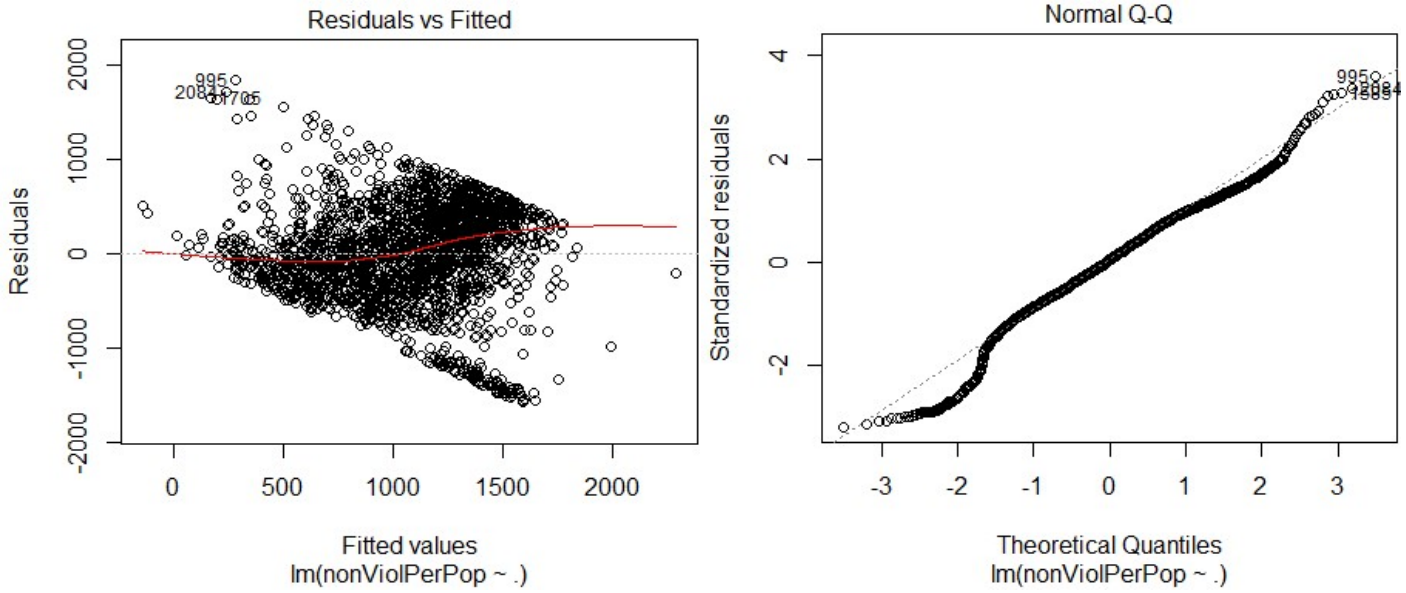
```
"nonViolPerPop"

"State"                 "pctSameHouse-5"        "pctSameCounty-5"       "pctSameState-5"
"perHoush"              "pctLargHous"           "pctPopDenseHous"       "pctSmallHousUnits"
"pctHousOccup"          "popDensity"
"pctLowEdu"             "pctNotHSgrad"          "pctCollGrad"
"pctBlack"              "pctWhite"              "pctAsian"              "pctHisp"
"pctSpeakOnlyEng"       "pctNotSpeakEng"
"pct16-24"              "pct65up"
"medIncome"             "pctWwage"              "pctWfarm"              "pctWdiv"
"pctWsocsec"            "pctPubAsst"            "pctRetire"             "medFamIncome"
"perCapInc"             "whitePerCap"           "blackPerCap"           "NAperCap"
"asianPerCap            "hispPerCap"            "pctPoverty"            "pctUnemploy"
"pctEmploy"             "pctEmployProfServ"     "pctOccupMgmt"          "pctHousOwnerOccup"
"pctVacantBoarded"      "pctHousWOphone"        "pctHousWOplumb"
"pctMaleNevMar"         "pctAllDivorc"          "pctKids2Par"
"pct12-17w2Par"         "pctWorkMom-18"         "pctKidsBornNevrMarr"
"pctImmig-3"            "pctImmig-5"            "pctImmig-8"            "pctImmig-10"
"pctForeignBorn"
"pctUsePubTrans"
```

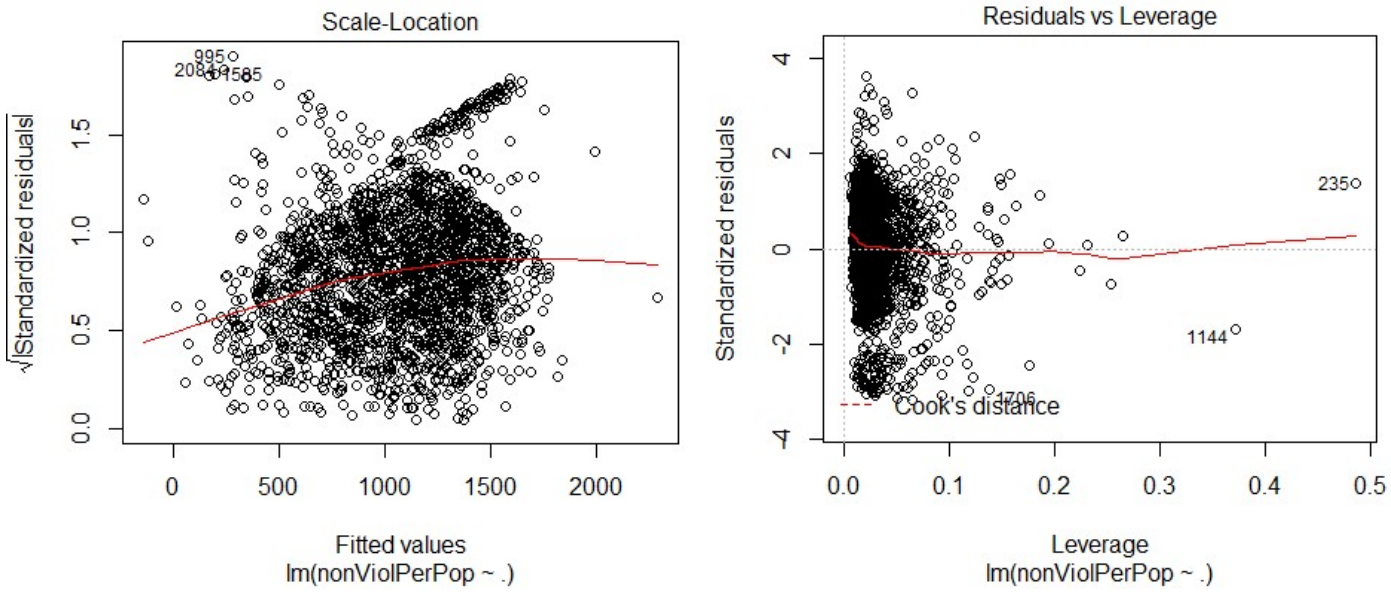
Legenda:

- Variabili riguardanti la posizione geografica
- Variabili riguardanti la densità di popolazione
- Variabili riguardanti l'educazione
- Variabili riguardanti l'etnia
- Variabili riguardanti l'età
- Variabili riguardanti il reddito
- Variabili riguardanti la situazione familiare
- Variabili riguardanti l'immigrazione

Modello iniziale con tutte le variabili rimaste.

```
summary(lm_start)

## Call:
## lm(formula = nonViolPerPop ~ ., data = covariate)
##
## Residual standard error: 511.4 on 2054 degrees of freedom
## Multiple R-squared:  0.3174, Adjusted R-squared:  0.2964
## F-statistic: 15.16 on 63 and 2054 DF,  p-value: < 2.2e-16
```





Ultimo modello stepwise selection.

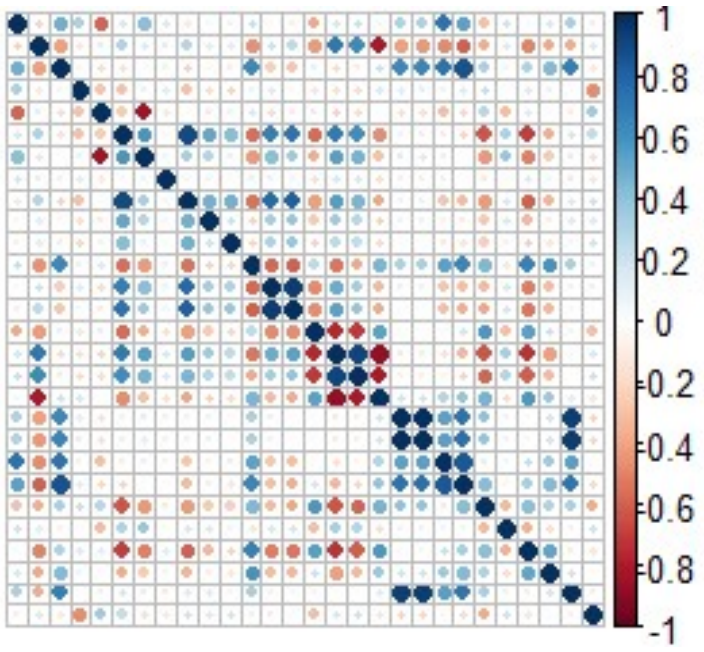
```
step(lm_start, direction="both")
```

```
## Step:  AIC=26443.28
## nonViolPerPop ~ State + perHoush + pctWhite + pctHisp + `pct16-24` +
##     pct65up + medIncome + pctWwage + pctWfarm + perCapInc + blackPerCap +
##     asianPerCap + pctLowEdu + pctCollGrad + pctOccupMgmt + pctAllDivorc +
##     pctKids2Par + `pct12-17w2Par` + pctKidsBornNevrMarr + `pctImmig-5` +
##     `pctImmig-8` + pctLargHous + pctPopDenseHous + pctSmallHousUnits +
##     pctHousOccup + pctHousWOphone + pctHousWOplumb + pctForeignBorn +
##     `pctSameHouse-5`

##
## Residual standard error: 509.7 on 2081 degrees of freedom
## Multiple R-squared:  0.313,  Adjusted R-squared:  0.3011
## F-statistic: 26.34 on 36 and 2081 DF,  p-value: < 2.2e-16
```

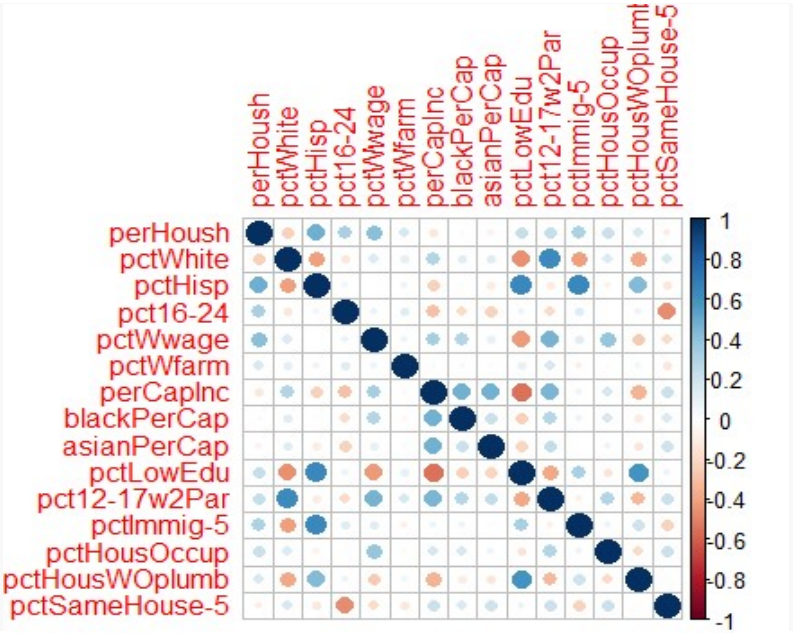
Multicollinearità: VIF e grafico correlazioni iniziale

```
imcdiag(X,y, method = 'TOL')
```

```
##                       TOL detection                          TOL detection
## perHoush           0.1030         0   ## pctWhite           0.1754         0
## pctHisp            0.1647         0   ## pct16-24           0.1812         0
## pct65up            0.0963         1   ## medIncome          0.0378         1
## pctWwage           0.1029         0   ## pctWfarm           0.7303         0
## perCapInc          0.0558         1   ## blackPerCap        0.6914         0
## asianPerCap        0.7420         0   ## pctLowEdu          0.1586         0
## pctCollGrad        0.0499         1   ## pctOccupMgmt       0.0597         1
## pctAllDivorc       0.1067         0   ## pctKids2Par        0.0317         1
## pct12-17w2Par      0.1404         0   ## pctKidsBornNevrMarr 0.1106         0
## pctImmig-5         0.0099         1   ## pctImmig-8         0.0072         1
## pctLargHous        0.1278         0   ## pctPopDenseHous    0.0622         1
## pctSmallHousUnits  0.1795         0   ## pctHousOccup       0.6423         0
## pctHousWOphone     0.2063         0   ## pctHousWOplumb     0.5680         0
## pctForeignBorn     0.0497         1   ## pctSameHouse-5     0.3305         0
```



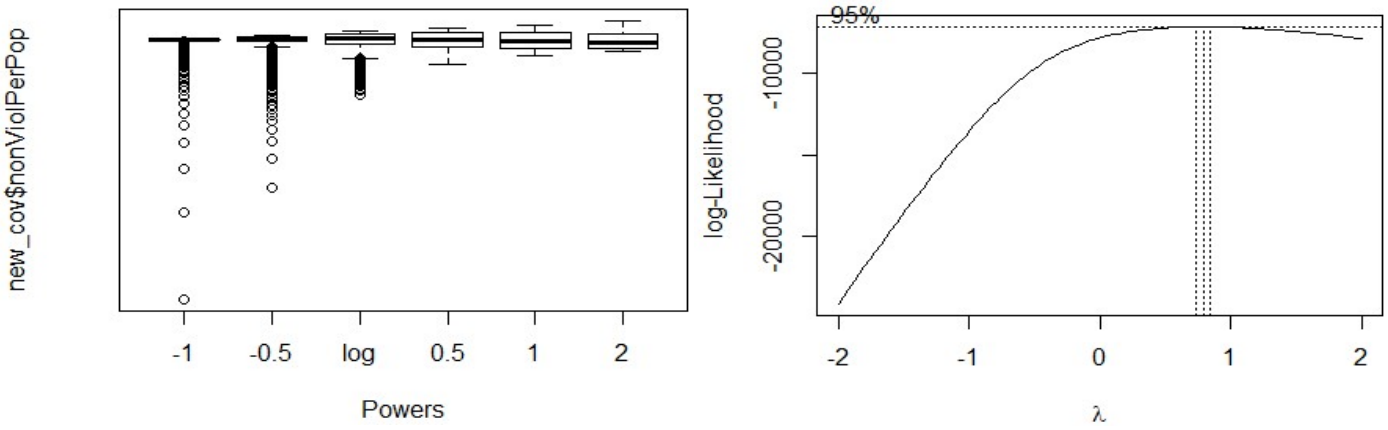
```
imcdiag(X,y)

##
##                    VIF    TOL
## perHoush        2.6400 0.3788
## pctWhite        2.8263 0.3538
## pctHisp         3.6080 0.2772
## pct16-24        1.7502 0.5714
## pctWwage        2.4228 0.4127
## pctWfarm        1.1911 0.8396
## perCapInc       2.7053 0.3696
## blackPerCap     1.3787 0.7253
## asianPerCap     1.3233 0.7557
## pctLowEdu       4.0036 0.2498
## pct12-17w2Par   3.4804 0.2873
## pctImmig-5      2.4172 0.4137
## pctHousOccup    1.4180 0.7052
## pctHousWOplumb 1.6530 0.6050
## pctSameHouse-5 1.9864 0.5034
```



Significatività variabili non collineari e riassunto modello.

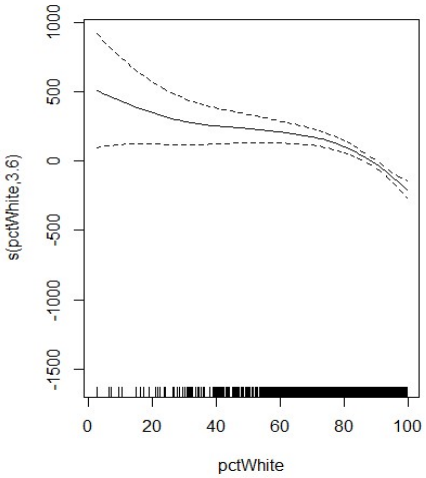
```
## Single term deletions

## Model:
## nonViolPerPop ~ State + perHoush + pctWhite + pctHisp + `pct16-24` +
##     pctWwage + pctWfarm + perCapInc + blackPerCap + asianPerCap +
##     pctLowEdu + `pct12-17w2Par` + `pctImmig-5` + pctHousOccup +
##     pctHousWOplumb + `pctSameHouse-5`


##                   Df Sum of Sq       RSS   AIC F value    Pr(>F)
## <none>                          568323519 26523
## State              8  16501456 584824975 26568  7.6000 4.620e-10 ***
## perHoush           1  21068095 589391614 26598 77.6258 < 2.2e-16 ***
## pctWhite           1   7882345 576205863 26550 29.0427 7.876e-08 ***
## pctHisp            1   1736682 570060201 26527  6.3988 0.0114923 *
## `pct16-24`         1   1025773 569349292 26525  3.7795 0.0520185 .
## pctWwage           1    180091 568503610 26522  0.6636 0.4154018
## pctWfarm           1   3095406 571418924 26532 11.4051 0.0007458 ***
## perCapInc          1   2002995 570326514 26528  7.3801 0.0066493 **
## blackPerCap        1   1453484 569777003 26526  5.3554 0.0207546 *
## asianPerCap        1    560912 568884431 26523  2.0667 0.1506971
## pctLowEdu          1   2753171 571076690 26531 10.1441 0.0014689 **
## `pct12-17w2Par`    1    968925 569292444 26525  3.5700 0.0589694 .
## `pctImmig-5`       1   1894208 570217727 26528  6.9792 0.0083074 **
## pctHousOccup       1    331311 568654830 26522  1.2207 0.2693458
## pctHousWOplumb     1    544388 568867907 26523  2.0058 0.1568463
## `pctSameHouse-5`   1   2467209 570790728 26530  9.0905 0.0026003 **
## ---

## Residual standard error: 521 on 2094 degrees of freedom
## Multiple R-squared:  0.2779, Adjusted R-squared:   0.27
## F-statistic: 35.04 on 23 and 2094 DF,  p-value: < 2.2e-16
```

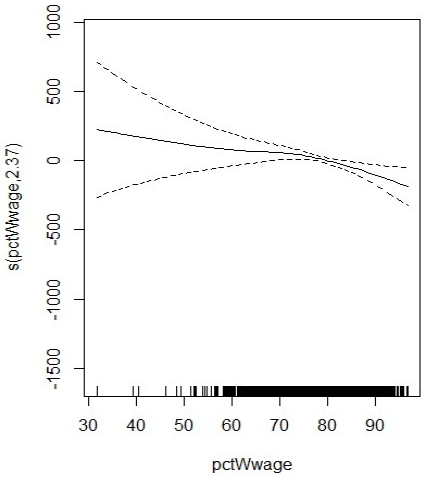
Symbox e Box-Cox transformation

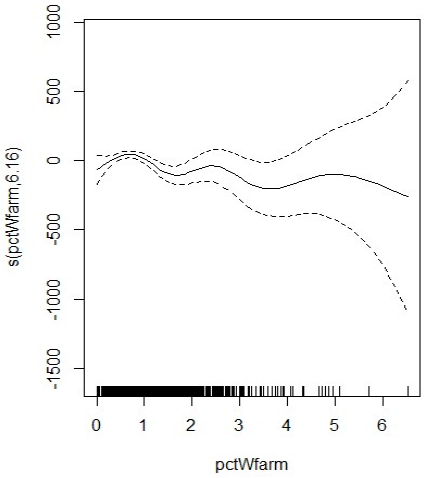


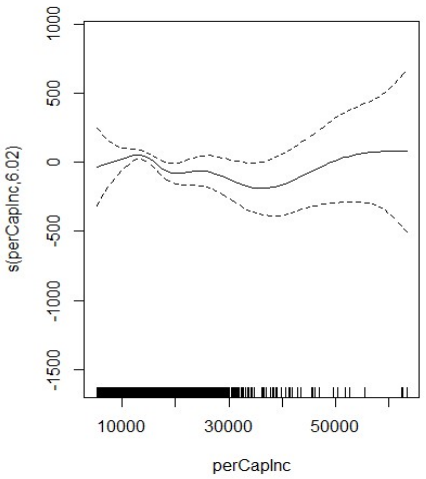
Gam(x)

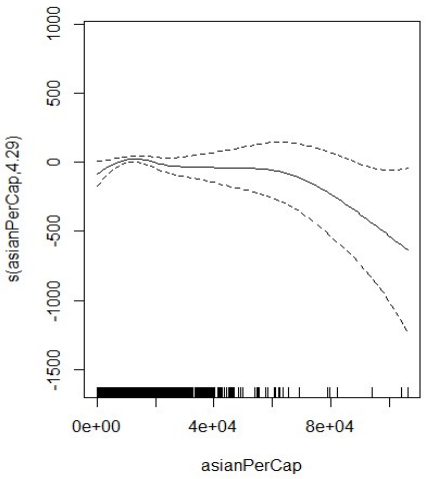
```
## Family: gaussian
## Link function: identity
## Formula:
## nonViolPerPop ~ State + s(perHoush) + s(pctWhite) + s(pctHisp) +
##     s(pctWwage) + s(pctWfarm) + s(perCapInc) + s(blackPerCap) +
##     s(asianPerCap) + s(pctLowEdu) + s(pctHousOccup) + s(pctHousWOplumb) +
##     s(pctYoung) + s(pctTeenW2Par) + s(pctImmig5) + s(pctSameHouse5)


## Approximate significance of smooth terms:
##                       edf Ref.df      F  p-value
## s(perHoush)         1.000  1.000 21.004 4.85e-06 ***
## s(pctWhite)         3.598  4.514 11.206 1.24e-09 ***
## s(pctHisp)          2.713  3.415  0.823  0.43378
## s(pctWwage)         2.372  3.085  2.820  0.03894 *
## s(pctWfarm)         6.163  7.276  2.649  0.00850 **
## s(perCapInc)        6.021  7.227  2.000  0.06605 .
## s(blackPerCap)      1.000  1.000  2.642  0.10421
## s(asianPerCap)      4.288  5.350  1.957  0.07964 .
## s(pctLowEdu)        1.000  1.000  5.159  0.02323 *
## s(pctHousOccup)     4.256  5.299  1.187  0.32303
## s(pctHousWOplumb)   1.000  1.000  1.077  0.29960
## s(pctYoung)         7.471  8.436  2.299  0.01275 *
## s(pctTeenW2Par)     4.473  5.601  7.517 1.65e-07 ***
## s(pctImmig5)        6.102  7.246  2.660  0.00891 **
## s(pctSameHouse5)    1. 000  1.000  3.254  0.07140 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.336   Deviance explained = 35.5%
## GCV = 2.5413e+05  Scale est. = 2.4676e+05  n = 2118
```

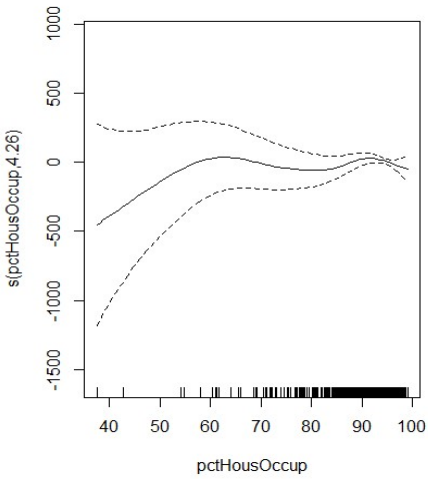

s(pctWhite,3.6)
1000
500
0
-500
-1000
-1500
0
20
40
60
80
100
pctWhite

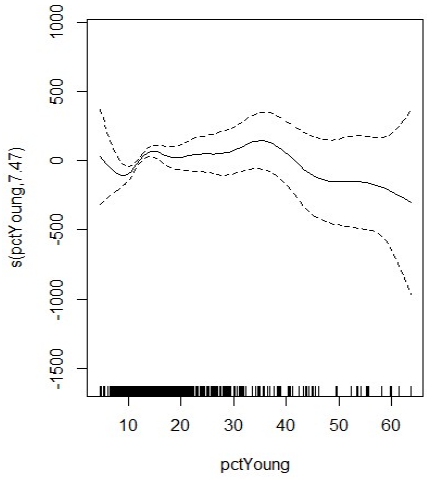

s(pctWwage,2.37)
1000
500
0
-500
-1000
-1500
30
40
50
60
70
80
90
pctWwage

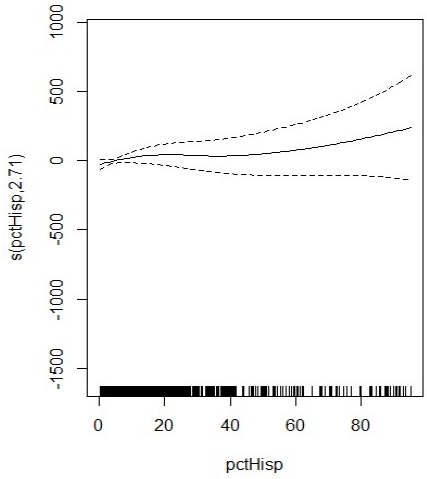

s(pctWfarm,6.16)
1000
500
0
-500
-1000
-1500
0
1
2
3
4
5
6
pctWfarm

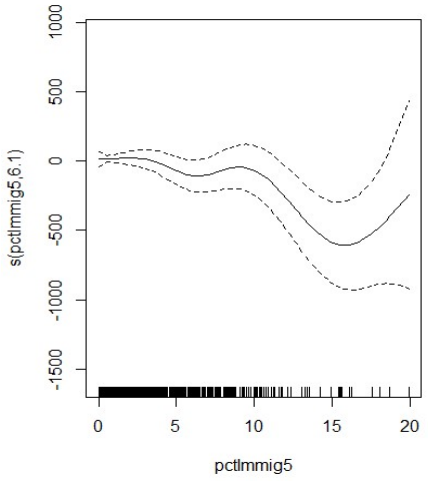

s(perCapInc,6.02)
1000
500
0
-500
-1000
-1500
10000
30000
50000
perCapInc

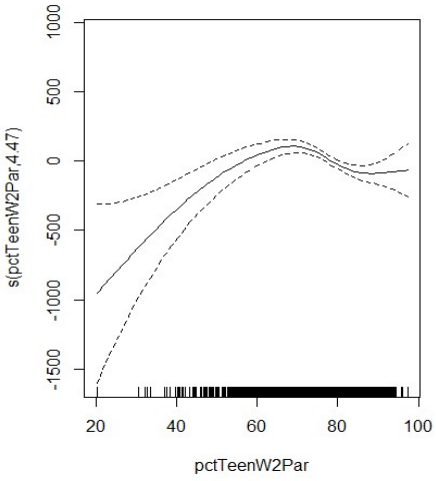

s(asianPerCap,4.29)
1000
500
0
-500
-1000
-1500
0e+00
4e+04
8e+04
asianPerCap

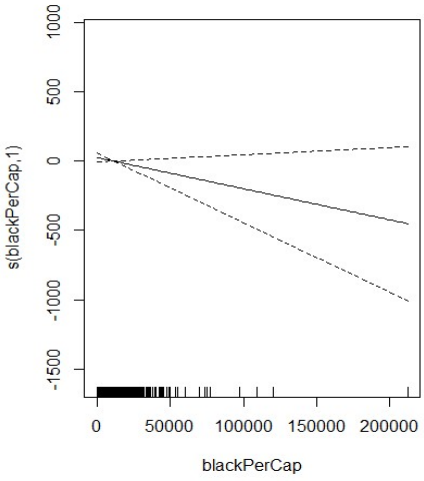

s(pctHousOccup,4.26)
1000
500
0
-500
-1000
-1500
40
50
60
70
80
90
100
pctHousOccup

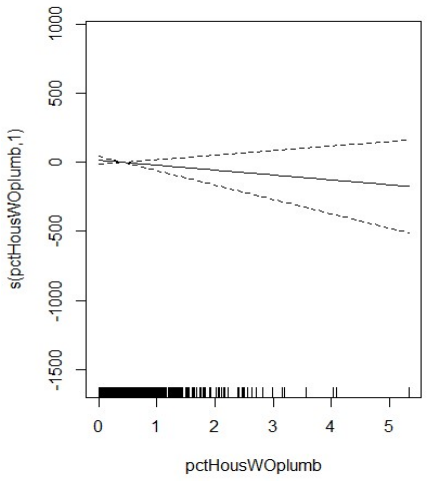

s(pctYoung,7.47)
1000
500
0
-500
-1000
-1500
10
20
30
40
50
60
pctYoung

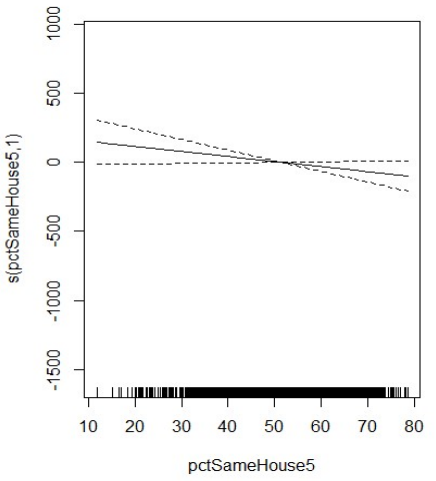

s(pctHisp,2.71)
1000
500
0
-500
-1000
-1500
0
20
40
60
80
pctHisp

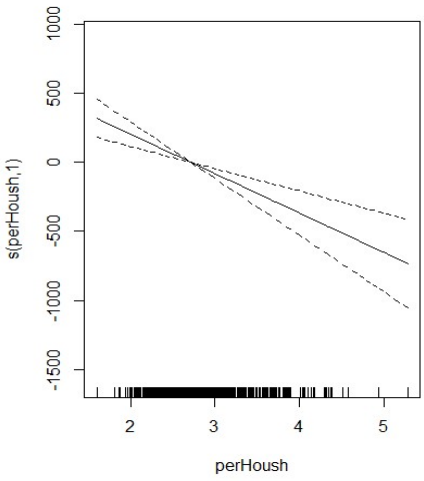

s(pctImmig5,6.1)
1000
500
0
-500
-1000
-1500
0
5
10
15
20
pctImmig5

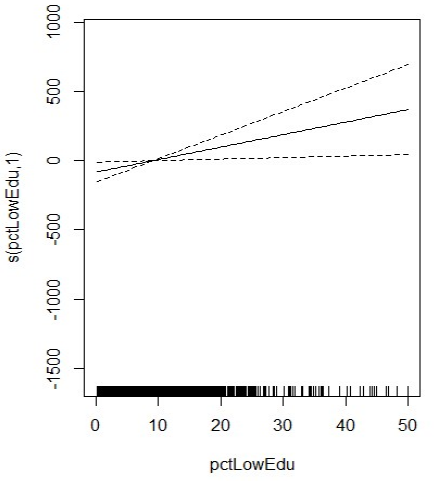

s(pctTeenW2Par,4.47)
1000
500
0
-500
-1000
-1500
20
40
60
80
100
pctTeenW2Par


s(blackPerCap,1)
1000
500
0
-500
-1000
-1500
0
50000
100000
150000
200000
blackPerCap


s(pctHousWOplumb,1)
1000
500
0
-500
-1000
-1500
0
1
2
3
4
5
pctHousWOplumb


s(pctSameHouse5,1)
1000
500
0
-500
-1000
-1500
10
20
30
40
50
60
70
80
pctSameHouse5


s(perHoush,1)
1000
500
0
-500
-1000
-1500
2
3
4
5
perHoush


s(pctLowEdu,1)
1000
500
0
-500
-1000
-1500
0
10
20
30
40
50
pctLowEdu

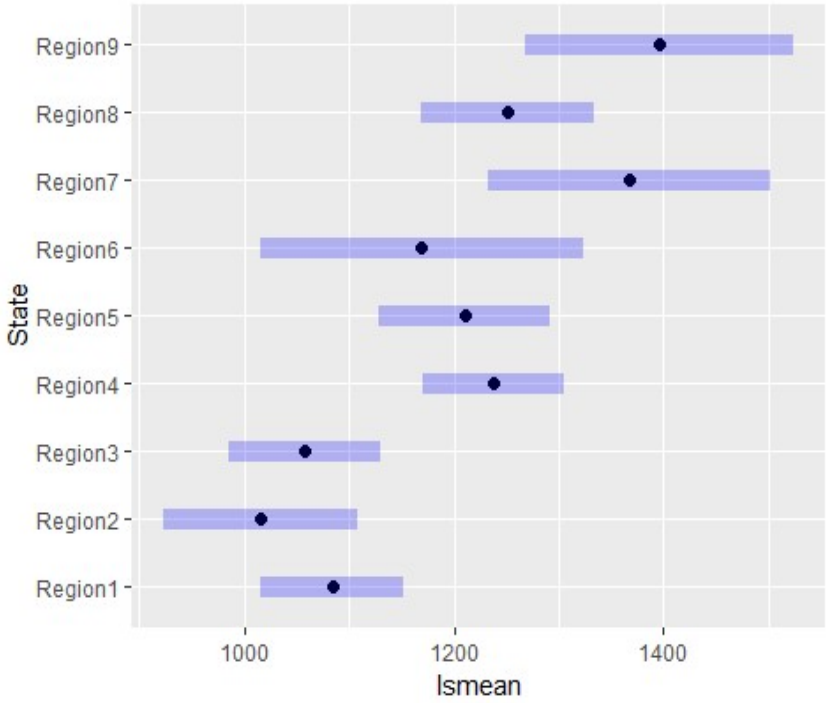
Confronto modelli con trasformazioni sulle x quantitative

```
anova(lm_2, lm_3, lm_4, lm_gam, test='Chisq')
```

```
## Analysis of Variance Table
##
## Model 1: nonViolPerPop ~ State + perHoush + pctWhite + pctHisp + `pct16-24` +
##     pctWwage + pctWfarm + perCapInc + blackPerCap + asianPerCap +
##     pctLowEdu + `pct12-17w2Par` + `pctImmig-5` + pctHousOccup +
##     pctHousWOplumb + `pctSameHouse-5`
## Model 2: nonViolPerPop ~ State + perHoush + pctWhite + I(pctWhite^2) +
##     pctHisp + pctWwage + +pctWfarm + perCapInc + blackPerCap +
##     log(asianPerCap + 1) + pctLowEdu + pctHousOccup + pctHousWOplumb +
##     pctYoung + log(pctTeenW2Par) + pctImmig5 + pctSameHouse5
## Model 3: nonViolPerPop ~ State + perHoush + pctWhite + I(pctWhite^2) +
##     pctHisp + pctWwage + pctWfarm + perCapInc + blackPerCap +
##     asianPerCap + I(asianPerCap^2) + pctLowEdu + pctHousOccup +
##     pctHousWOplumb + pctYoung + pctTeenW2Par + I(pctTeenW2Par^2) +
##     pctImmig5 + pctSameHouse5
## Model 4: nonViolPerPop ~ State + s(perHoush) + s(pctWhite) + s(pctHisp) +
##     s(pctWwage) + s(pctWfarm) + s(perCapInc) + s(blackPerCap) +
##     s(asianPerCap) + s(pctLowEdu) + s(pctHousOccup) + s(pctHousWOplumb) +
##     s(pctYoung) + s(pctTeenW2Par) + s(pctImmig5) + s(pctSameHouse5)


##   Res.Df       RSS     Df Sum of Sq  Pr(>Chi)
## 1 2094.0 568323519
## 2 2093.0 553657897  1.000  14665622 1.265e-14 ***
## 3 2091.0 541627623  2.000  12030274 2.590e-11 ***
## 4 2056.5 507462847 34.457  34164776 2.060e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tentativo di trasformazione della variabile categoriale State

```
## Region1 Region2 Region3 Region4 Region5 Region6 Region7 Region8 Region9
##     491     158     296     426     247      45      72     303      80
```




```
## Analysis of Variance Table
##
## Model 1: nonViolPerPop ~ State + perHoush + pctWhite + I(pctWhite^2) +
##     pctHisp + pctWwage + pctWfarm + perCapInc + blackPerCap +
##     asianPerCap + I(asianPerCap^2) + pctLowEdu + pctHousOccup +
##     pctHousWOplumb + pctYoung + pctTeenW2Par + I(pctTeenW2Par^2) +
##     pctImmig5 + pctSameHouse5
## Model 2: nonViolPerPop ~ State + perHoush + pctWhite + I(pctWhite^2) +
##     pctHisp + pctWwage + pctWfarm + perCapInc + blackPerCap +
##     asianPerCap + I(asianPerCap^2) + pctLowEdu + pctHousOccup +
##     pctHousWOplumb + pctYoung + pctTeenW2Par + I(pctTeenW2Par^2) +
##     pctImmig5 + pctSameHouse5
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   2091 541627623
## 2   2093 541716298 -2    -88675 0.1712 0.8427
```

Riassunto modello scelto (lm_4)

```
## Single term deletions
##
## Model:
## nonViolPerPop ~ State + perHoush + pctWhite + I(pctWhite^2) +
##     pctHisp + pctWwage + pctWfarm + perCapInc + blackPerCap +
##     asianPerCap + I(asianPerCap^2) + pctLowEdu + pctHousOccup +
##     pctHousWOplumb + pctYoung + pctTeenW2Par + I(pctTeenW2Par^2) +
##     pctImmig5 + pctSameHouse5
##                   Df Sum of Sq       RSS   AIC F value    Pr(>F)
## <none>                          541627623 26427
## State              8  14520920 556148543 26467  7.0074 3.691e-09 ***
## perHoush           1  10611997 552239620 26466 40.9685 1.904e-10 ***
## pctWhite           1   2129215 543756838 26433  8.2200 0.0041846 **
## I(pctWhite^2)      1   6239389 547867012 26449 24.0877 9.916e-07 ***
## pctHisp            1    426198 542053821 26427  1.6454 0.1997318
## pctWwage           1    692279 542319902 26428  2.6726 0.1022395
## pctWfarm           1   2206562 543834185 26434  8.5186 0.0035529 **
## perCapInc          1    766105 542393728 26428  2.9576 0.0856220 .
## blackPerCap        1   1011416 542639039 26429  3.9047 0.0482837 *
## asianPerCap        1    154766 541782389 26426  0.5975 0.4396257
## I(asianPerCap^2)   1    756717 542384340 26428  2.9214 0.0875620 .
## pctLowEdu          1   2332244 543959868 26434  9.0038 0.0027260 **
## pctHousOccup       1    127562 541755185 26426  0.4925 0.4829076
## pctHousWOplumb     1    231054 541858677 26426  0.8920 0.3450439
## pctYoung           1    351264 541978887 26426  1.3561 0.2443498
## pctTeenW2Par       1   9882425 551510049 26463 38.1520 7.842e-10 ***
## I(pctTeenW2Par^2)  1  11097339 552724962 26468 42.8422 7.442e-11 ***
## pctImmig5          1   3202217 544829840 26438 12.3624 0.0004474 ***
## pctSameHouse5      1   1776775 543404398 26432  6.8594 0.0088814 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 508.9 on 2091 degrees of freedom
## Multiple R-squared:  0.3118, Adjusted R-squared:  0.3033
## F-statistic: 36.44 on 26 and 2091 DF,  p-value: < 2.2e-16
```

Test White per Eteroschedasticità e standard error calcolati con la correzione di White

```
ncvTest(lm_4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 49.74261, Df = 1, p = 1.753e-12
```

```
coeftest(lm_4, vcov. = vcovHC(lm_4))
```

```
##
## t test of coefficients:
##
##                       Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)        6.3549e+01  6.1505e+02  0.1033 0.9177165
## StateRegion2      -6.8843e+01  4.6560e+01 -1.4786 0.1394039
## StateRegion3      -2.6733e+01  5.6243e+01 -0.4753 0.6346164
## StateRegion4       1.5309e+02  3.6195e+01  4.2296 2.442e-05 ***
## StateRegion5       1.2586e+02  5.4039e+01  2.3290 0.0199555 *
## StateRegion6       8.4935e+01  8.6502e+01  0.9819 0.3262704
## StateRegion7       2.8327e+02  7.1952e+01  3.9369 8.524e-05 ***
## StateRegion8       1.6648e+02  5.6807e+01  2.9306 0.0034196 **
## StateRegion9       3.1225e+02  7.9210e+01  3.9421 8.344e-05 ***
## perHoush          -3.5919e+02  5.5122e+01 -6.5163 9.006e-11 ***
## pctWhite           1.4240e+01  7.0388e+00  2.0232 0.0431841 *
## I(pctWhite^2)     -1.6946e-01  4.7089e-02 -3.5988 0.0003272 ***
## pctHisp            2.0425e+00  1.8422e+00  1.1088 0.2676600
## pctWwage          -3.8324e+00  2.5900e+00 -1.4797 0.1391022
## pctWfarm          -5.8033e+01  2.1864e+01 -2.6542 0.0080091 **
## perCapInc         -5.2729e-03  3.0541e-03 -1.7265 0.0844105 .
## blackPerCap       -2.7557e-03  1.7002e-03 -1.6208 0.1052071
## asianPerCap        2.1747e-03  3.0419e-03  0.7149 0.4747436
## I(asianPerCap^2)  -7.0132e-08  4.8044e-08 -1.4597 0.1445148
## pctLowEdu          1.0802e+01  4.0900e+00  2.6412 0.0083237 **
## pctHousOccup       1.9893e+00  3.4000e+00  0.5851 0.5585430
## pctHousWOplumb    -3.2159e+01  4.2969e+01 -0.7484 0.4542887
## pctYoung           3.0118e+00  2.6875e+00  1.1207 0.2625623
## pctTeenW2Par       7.2448e+01  1.5489e+01  4.6774 3.091e-06 ***
## I(pctTeenW2Par^2) -5.3782e-01  1.0400e-01 -5.1714 2.545e-07 ***
## pctImmig5         -2.6235e+01  8.2603e+00 -3.1760 0.0015148 **
## pctSameHouse5     -4.7249e+00  1.8318e+00 -2.5793 0.0099671 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modello dopo aver eliminato una a una la variabile meno significativa

```
ncvTest(lm_5_7)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 52.76469, Df = 1, p = 3.76e-13
```

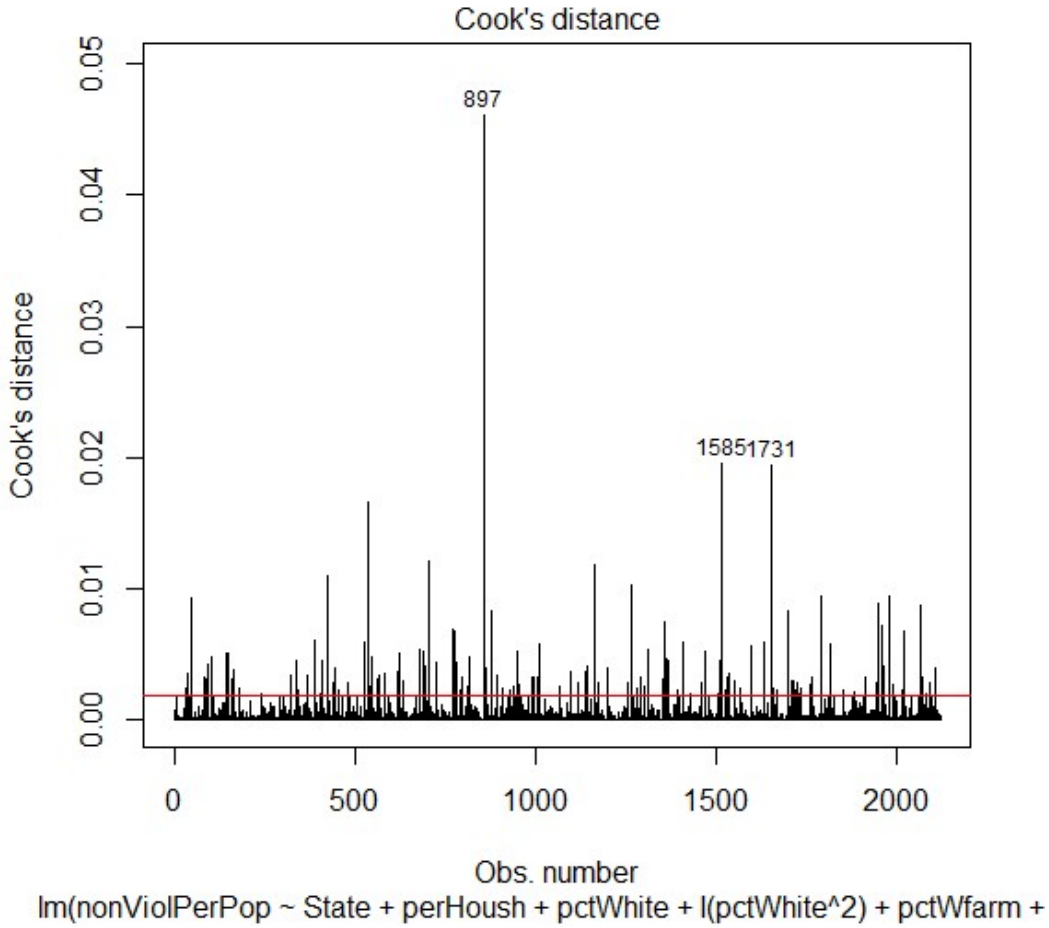
```
coeftest(lm_5_7, vcov. = vcovHC(lm_5_7))
```

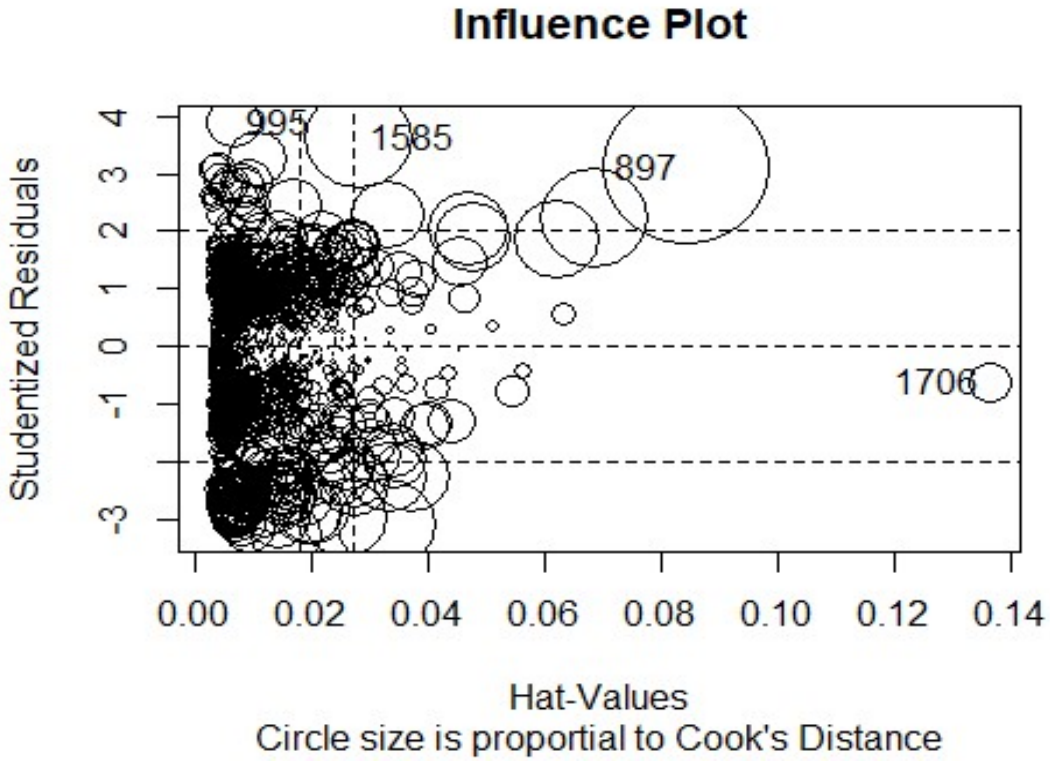
```
##
## t test of coefficients:
##
##                      Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)       -4.1783e+01  5.8125e+02 -0.0719 0.9426997
## StateRegion2      -5.9645e+01  4.5013e+01 -1.3251 0.1852947
## StateRegion3      -2.0682e+01  5.4363e+01 -0.3804 0.7036604
## StateRegion4       1.6049e+02  3.3121e+01  4.8456 1.354e-06 ***
## StateRegion5       1.3815e+02  5.0888e+01  2.7148 0.0066855 **
## StateRegion6       1.0031e+02  8.5956e+01  1.1669 0.2433657
## StateRegion7       2.9868e+02  6.7983e+01  4.3935 1.171e-05 ***
## StateRegion8       1.9489e+02  4.5278e+01  4.3044 1.752e-05 ***
## StateRegion9       3.3199e+02  7.3652e+01  4.5076 6.918e-06 ***
## perHoush          -3.5360e+02  4.3654e+01 -8.1001 9.214e-16 ***
## pctWhite           1.5896e+01  6.9941e+00  2.2728 0.0231385 *
## I(pctWhite^2)     -1.7659e-01  4.6865e-02 -3.7679 0.0001691 ***
## pctWfarm          -5.8526e+01  2.1519e+01 -2.7198 0.0065869 **
## perCapInc         -8.3579e-03  2.7614e-03 -3.0267 0.0025026 **
## pctLowEdu          1.2927e+01  2.9308e+00  4.4109 1.082e-05 ***
## pctTeenW2Par       7.1625e+01  1.5464e+01  4.6318 3.848e-06 ***
## I(pctTeenW2Par^2) -5.4122e-01  1.0419e-01 -5.1946 2.250e-07 ***
## pctImmig5         -2.1525e+01  7.7885e+00 -2.7637 0.0057645 **
## pctSameHouse5     -4.6506e+00  1.4728e+00 -3.1576 0.0016133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_5_7)
```

```
##
## Residual standard error: 510 on 2099 degrees of freedom
## Multiple R-squared:  0.3064, Adjusted R-squared:  0.3004
## F-statistic: 51.51 on 18 and 2099 DF,  p-value: < 2.2e-16
```

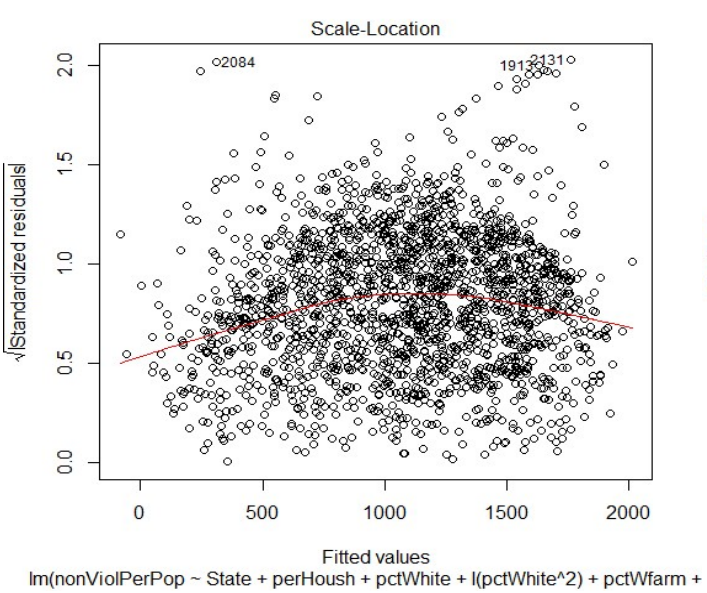
Grafico delle distanze di Cook e Influence plot del modello con tutte covariate significative

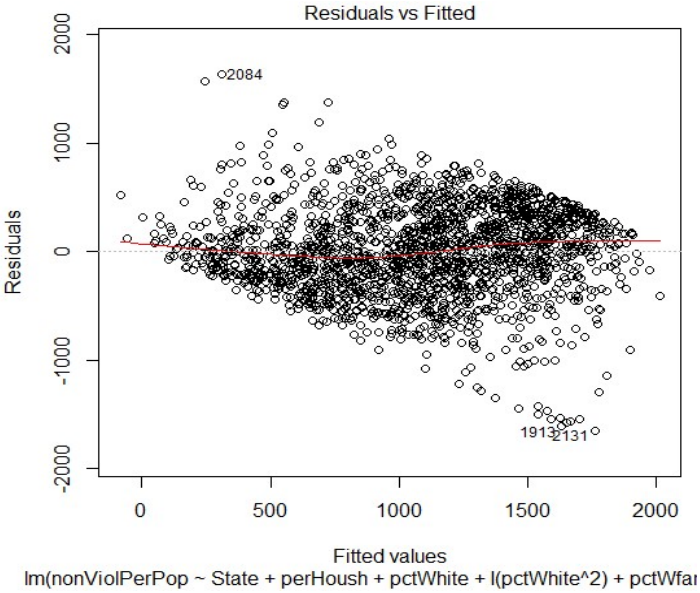


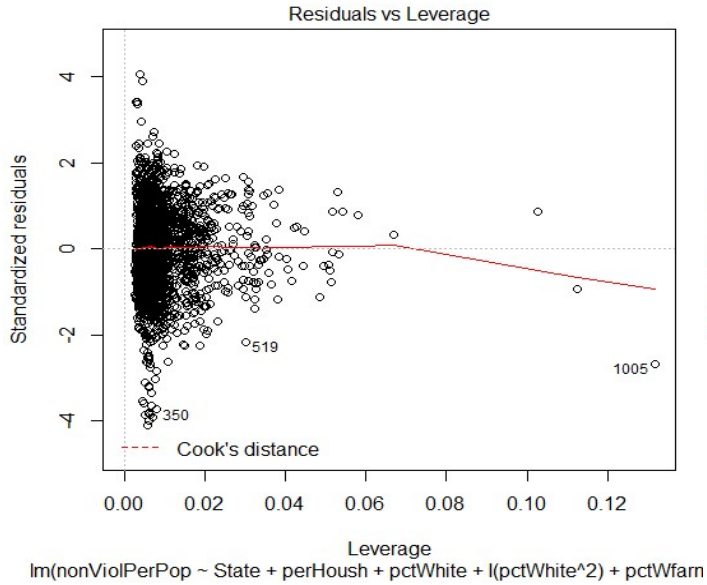


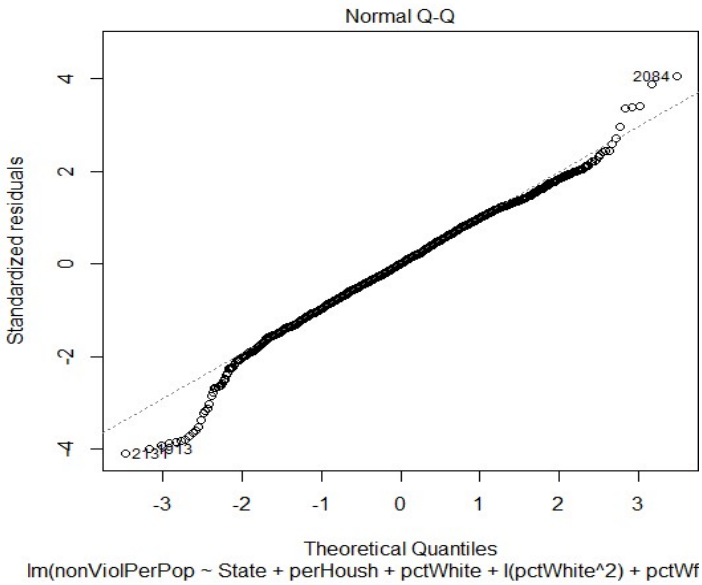
Modello finale con grafici di diagnostica.

```
## Model:
## nonViolPerPop ~ State + perHoush + pctWhite + I(pctWhite^2) +
##     pctWfarm + perCapInc + pctLowEdu + pctTeenW2Par + I(pctTeenW2Par^2) +
##     pctImmig5 + pctSameHouse5

## Residual standard error: 403.5 on 1948 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.5165
## F-statistic: 117.7 on 18 and 1948 DF,  p-value: < 2.2e-16
```



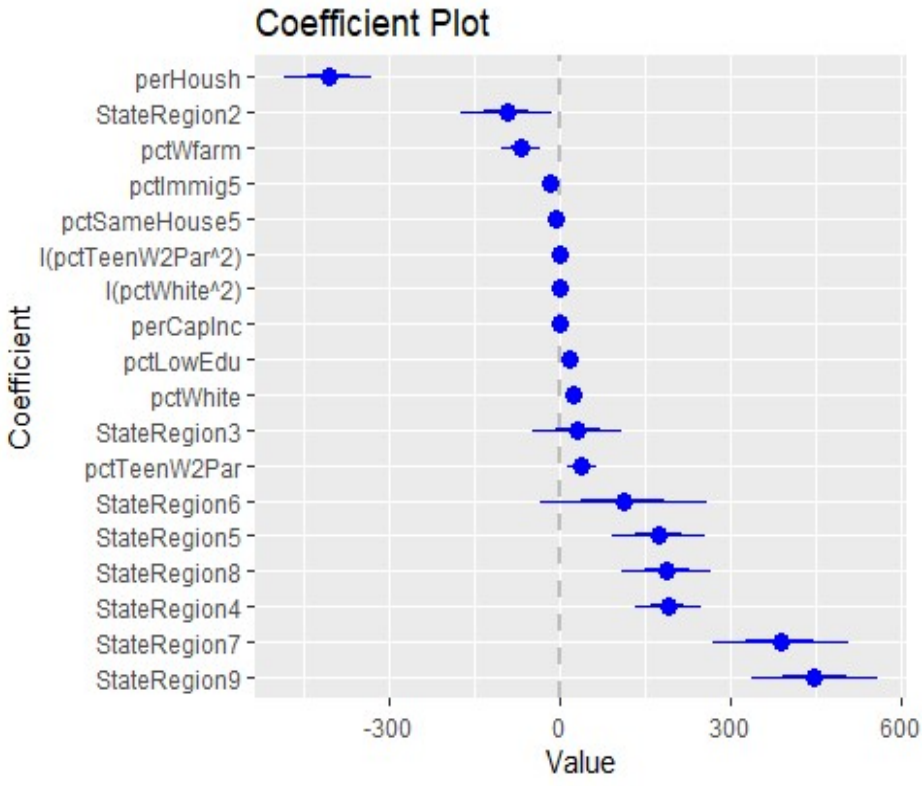






Test e correzione di White per Eteroschedasticità sul modello senza punti influenti

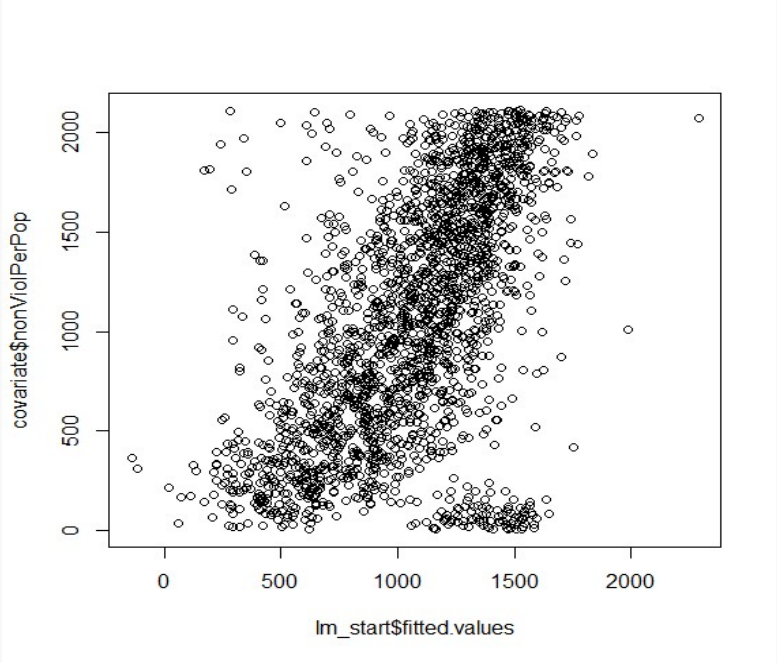
```
ncvTest(lm_noInf)
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 12.27118, Df = 1, p = 0.00046001
coeftest(lm_noInf, vcov. = vcovHC(lm_noInf))
## t test of coefficients:
##
##                       Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)         1.2679e+03  6.3396e+02   1.9999  0.045648 *
## StateRegion2       -9.2558e+01  3.6750e+01  -2.5186  0.011863 *
## StateRegion3        3.1545e+01  4.4246e+01   0.7129  0.475976
## StateRegion4        1.8984e+02  2.8623e+01   6.6325 4.262e-11 ***
## StateRegion5        1.7335e+02  3.9399e+01   4.3999 1.142e-05 ***
## StateRegion6        1.1228e+02  6.2440e+01   1.7982  0.072301 .
## StateRegion7        3.8730e+02  5.1548e+01   7.5134 8.727e-14 ***
## StateRegion8        1.8800e+02  3.6863e+01   5.1001 3.724e-07 ***
## StateRegion9        4.4672e+02  5.5332e+01   8.0734 1.184e-15 ***
## perHoush           -4.0521e+02  3.6567e+01 -11.0812 < 2.2e-16 ***
## pctWhite            2.5053e+01  4.4634e+00   5.6130 2.274e-08 ***
## I(pctWhite^2)      -2.5101e-01  3.1267e-02  -8.0280 1.696e-15 ***
## pctWfarm           -6.8273e+01  1.5102e+01  -4.5208 6.530e-06 ***
## perCapInc          -9.4862e-03  2.2030e-03  -4.3061 1.744e-05 ***
## pctLowEdu           1.5728e+01  2.2329e+00   7.0437 2.587e-12 ***
## pctTeenW2Par        3.9159e+01  1.6364e+01   2.3929  0.016809 *
## I(pctTeenW2Par^2) -3.5321e-01  1.0963e-01  -3.2217  0.001295 **
## pctImmig5          -1.6602e+01  5.6637e+00  -2.9313  0.003414 **
## pctSameHouse5      -5.0755e+00  1.2428e+00  -4.0839 4.607e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



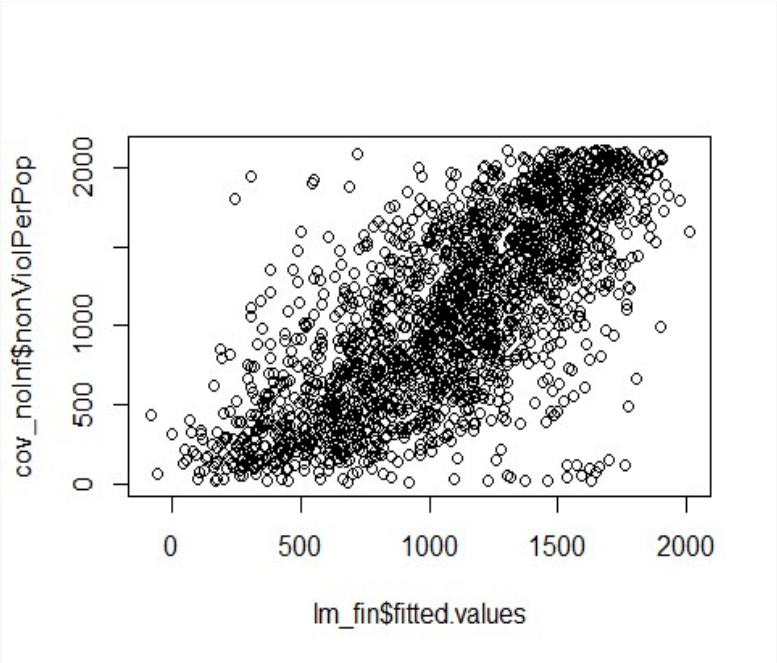
Contrasti tra i livelli di State

```
##  contrast             estimate       SE   df t.ratio p.value
##  Region1 - Region2   92.557817 39.68652 1948   2.332  0.3231
##  Region1 - Region3  -31.544510 39.06264 1948  -0.808  0.9967
##  Region1 - Region4 -189.842198 29.22488 1948  -6.496  <.0001
##  Region1 - Region5 -173.352498 41.25174 1948  -4.202  0.0009
##  Region1 - Region6 -112.279642 73.25831 1948  -1.533  0.8402
##  Region1 - Region7 -387.304804 59.56623 1948  -6.502  <.0001
##  Region1 - Region8 -188.003533 38.68469 1948  -4.860  <.0001
##  Region1 - Region9 -446.718732 55.61611 1948  -8.032  <.0001
##  Region2 - Region3 -124.102327 44.49666 1948  -2.789  0.1192
##  Region2 - Region4 -282.400014 39.13361 1948  -7.216  <.0001
##  Region2 - Region5 -265.910314 47.08046 1948  -5.648  <.0001
##  Region2 - Region6 -204.837458 77.29973 1948  -2.650  0.1669
##  Region2 - Region7 -479.862621 64.81537 1948  -7.404  <.0001
##  Region2 - Region8 -280.561350 48.22525 1948  -5.818  <.0001
##  Region2 - Region9 -539.276549 60.62273 1948  -8.896  <.0001
##  Region3 - Region4 -158.297688 36.77420 1948  -4.305  0.0006
##  Region3 - Region5 -141.807987 39.51330 1948  -3.589  0.0103
##  Region3 - Region6  -80.735132 74.95924 1948  -1.077  0.9775
##  Region3 - Region7 -355.760294 62.42813 1948  -5.699  <.0001
##  Region3 - Region8 -156.459023 42.24109 1948  -3.704  0.0068
##  Region3 - Region9 -415.174222 56.59526 1948  -7.336  <.0001
##  Region4 - Region5   16.489700 39.22675 1948   0.420  1.0000
##  Region4 - Region6   77.562556 72.43776 1948   1.071  0.9783
##  Region4 - Region7 -197.462606 57.63212 1948  -3.426  0.0181
##  Region4 - Region8    1.838664 39.73749 1948   0.046  1.0000
##  Region4 - Region9 -256.876534 53.65318 1948  -4.788  0.0001
##  Region5 - Region6   61.072856 75.67325 1948   0.807  0.9967
##  Region5 - Region7 -213.952307 60.31905 1948  -3.547  0.0119
##  Region5 - Region8  -14.651036 42.65094 1948  -0.344  1.0000
##  Region5 - Region9 -273.366235 58.13181 1948  -4.703  0.0001
##  Region6 - Region7 -275.025162 87.26819 1948  -3.151  0.0435
##  Region6 - Region8  -75.723891 76.99144 1948  -0.984  0.9874
##  Region6 - Region9 -334.439090 85.01547 1948  -3.934  0.0028
##  Region7 - Region8  199.301271 62.73210 1948   3.177  0.0403
##  Region7 - Region9  -59.413928 71.57476 1948  -0.830  0.9960
##  Region8 - Region9 -258.715199 57.15804 1948  -4.526  0.0002
##
## P value adjustment: tukey method for comparing a family of 9 estimates
```

Grafico previsti vs osservati con rispettiva correlazione e R2 aggiustato del modello finale




```
## Residual standard error: 511.4 on 2054 degrees of freedom
## Multiple R-squared:  0.3174, Adjusted R-squared:  0.2964
## F-statistic: 15.16 on 63 and 2054 DF,  p-value: < 2.2e-16

## "cor(previsti-osservati modello iniziale)=0.563364"
```




```
## Residual standard error: 403.5 on 1948 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.5165
## F-statistic: 117.7 on 18 and 1948 DF,  p-value: < 2.2e-16

## "cor(previsti-osservati modello finale)=0.7217361"
```

Modello Logistico

Indici di posizione del target quantitativo e distribuzione del target dummy

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0   530.2  1057.5  1057.7  1584.8  2114.0

##
##   0   1
## 0.5 0.5
```

Primo modello

```
## Call:
## glm(formula = target_dummy ~ State + pctWhite + IncMil, family = binomial,
##     data = covariate)

## “Deviance first model=2579.854”

## “Null.Deviance=2936.171”

## “R2 first model=0.1213543”
```

LRT primo modello

```
## Single term deletions
##
## Model:
## target_dummy ~ State + pctWhite + IncMil
##          Df Deviance    AIC    LRT  Pr(>Chi)
## <none>        2579.8 2601.8
## State     8   2636.6 2642.6 56.787 1.982e-09 ***
## pctWhite  1   2629.4 2649.4 49.556 1.928e-12 ***
## IncMil    1   2660.6 2680.6 80.748 < 2.2e-16 ***
```

Odds Ratio del primo modello

```
##                 OR 2.5 % 97.5 %
## (Intercept)  18.83  9.61  37.75
## StateRegion2  1.02  0.69   1.51
## StateRegion3  2.01  1.44   2.82
## StateRegion4  1.66  1.24   2.22
## StateRegion5  2.09  1.47   2.99
## StateRegion6  1.99  1.05   3.82
## StateRegion7  1.31  0.77   2.20
## StateRegion8  2.57  1.84   3.61
## StateRegion9  3.46  2.07   5.92
## pctWhite      0.98  0.97   0.98
## IncMil        0.92  0.90   0.94
```

Secondo modello

```
## Call:
## glm(formula = target_dummy ~ pctWhite + IncMil + perHoush, family = binomial,
##     data = covariate)

## “Deviance first model=2514.44”

## “R2 first model=0.1436331”
```

LRT secondo modello

```
## Single term deletions
##
## Model:
## target_dummy ~ pctWhite + IncMil + perHoush
##          Df Deviance    AIC    LRT  Pr(>Chi)
## <none>        2514.4 2522.4
## pctWhite  1   2654.9 2660.9 140.43 < 2.2e-16 ***
## IncMil    1   2652.4 2658.4 137.96 < 2.2e-16 ***
## perHoush  1   2636.6 2642.6 122.20 < 2.2e-16 ***
## ---
```

Odds Ratio del secondo modello

```
##                    OR   2.5 %   97.5 %
## (Intercept) 14403.77 4279.35 50558.40
## pctWhite        0.96    0.95     0.97
## IncMil          0.90    0.88     0.92
## perHoush        0.19    0.14     0.26
```

Modello con quattro covariate

```
## Call:
## glm(formula = target_dummy ~ State + pctWhite + perHoush + IncMil,
##     family = binomial, data = covariate)

## “Deviance=2441.491”

## “R2=0.1684779”
```

LRT secondo modello

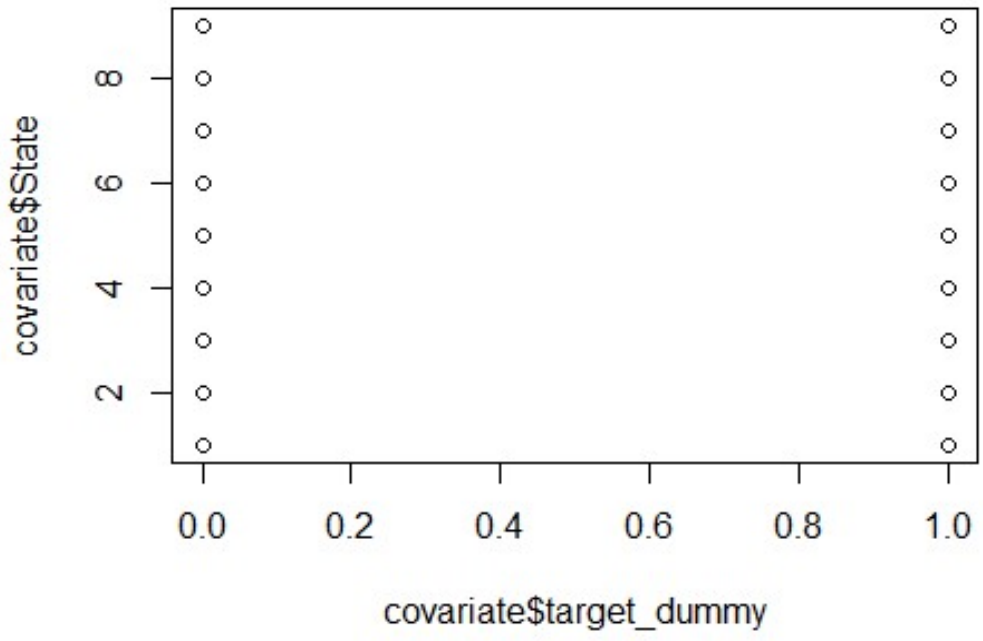
```
## Single term deletions
##
## Model:
## target_dummy ~ State + pctWhite + perHoush + IncMil
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>        2441.5 2465.5
## State     8   2514.4 2522.4  72.949 1.269e-12 ***
## pctWhite  1   2521.4 2543.4  79.886 < 2.2e-16 ***
## perHoush  1   2579.8 2601.8 138.363 < 2.2e-16 ***
## IncMil    1   2558.6 2580.6 117.094 < 2.2e-16 ***
## ---
```

OR del modello con quattro covariate

```
##                    OR   2.5 %   97.5 %
## (Intercept)  11790.33 3045.67 47709.75
## StateRegion2     0.74    0.49     1.13
## StateRegion3     1.25    0.88     1.79
## StateRegion4     1.49    1.10     2.02
## StateRegion5     2.11    1.46     3.05
## StateRegion6     1.51    0.78     2.95
## StateRegion7     1.46    0.83     2.54
## StateRegion8     3.33    2.35     4.76
## StateRegion9     2.57    1.51     4.45
## pctWhite         0.97    0.96     0.97
## perHoush         0.14    0.10     0.20
## IncMil           0.89    0.87     0.91
```

Valutiamo la separazione nel modello con quattro covariate

```
##             0   1
##   Region1 344 147
##   Region2  97  61
##   Region3 108 188
##   Region4 226 200
##   Region5  85 162
##   Region6  21  24
##   Region7  39  33
##   Region8 113 190
##   Region9  26  54
```



Confrontiamo le statistiche delle covariate quantitative nei due gruppi definiti dal target

```
summary(covariate_tar0)
```

```
##     pctWhite        perHoush         IncMil
##  Min.   : 7.26   Min.   :1.860   Min.   : 5.237
##  1st Qu.:86.36   1st Qu.:2.540   1st Qu.:12.717
##  Median :94.20   Median :2.720   Median :16.143
##  Mean   :88.59   Mean   :2.757   Mean   :17.511
##  3rd Qu.:97.44   3rd Qu.:2.905   3rd Qu.:20.172
##  Max.   :99.63   Max.   :5.280   Max.   :63.302
```

```
summary(covariate_tar1)
```

```
##     pctWhite        perHoush         IncMil
##  Min.   : 2.68   Min.   :1.600   Min.   : 5.561
##  1st Qu.:69.88   1st Qu.:2.480   1st Qu.:11.096
##  Median :83.17   Median :2.620   Median :12.822
##  Mean   :79.33   Mean   :2.668   Mean   :13.848
##  3rd Qu.:93.16   3rd Qu.:2.785   3rd Qu.:15.453
##  Max.   :99.22   Max.   :4.380   Max.   :62.376
```

Abbiamo classificato le unità secondo le probabilità previste secondo il modello logistico

```
##      target_dummy   State pctWhite perHoush IncMil predicted_p predicted_y
## 2210            0 Region1    98.48     2.57 16.201   0.3282763           0
## 2211            1 Region8    61.68     3.07 10.237   0.8029041           1
## 2212            1 Region5    76.65     2.68  9.995   0.7743833           1
## 2213            1 Region8    92.62     2.46 14.131   0.7555082           1
## 2214            1 Region5    69.91     2.89  8.100   0.7789779           1
## 2215            1 Region8    71.27     2.61 11.510   0.8628614           1
```

Matrice di confusione

```
##          predicted
## observed   0   1
##        0 734 325
##        1 265 794
```

```
##          predicted
## observed         0         1
##        0 0.3465534 0.1534466
##        1 0.1251180 0.374882
```

```
## “Accuracy=0.7214353”
```

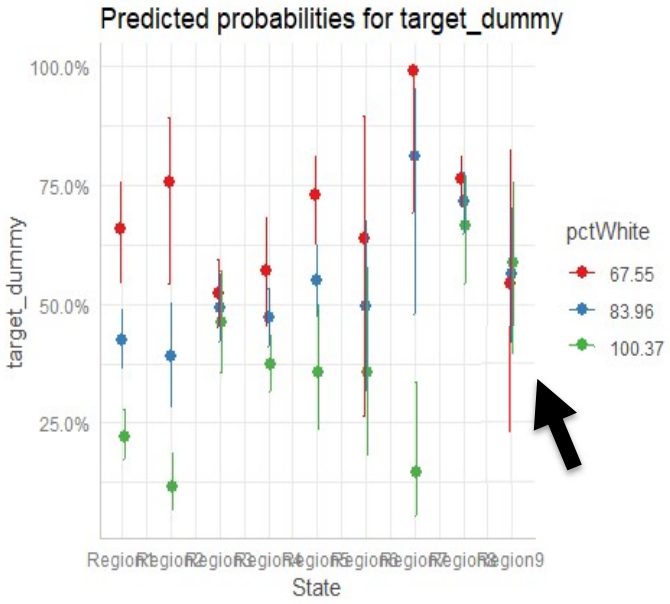
```
## “Error rate=0.2785647”
```

Abbiamo provato diversi modelli con interazioni, questo è quello migliore

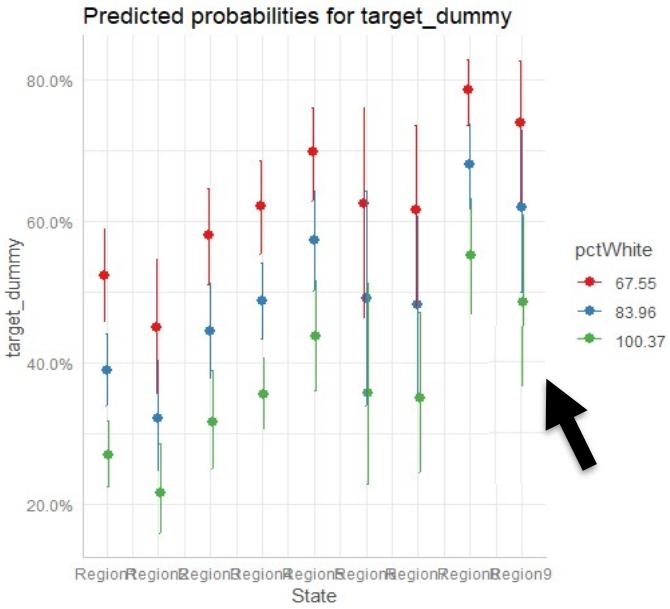
```
## Single term deletions
##
## Model:
## target_dummy ~ State * pctWhite + perHoush + IncMil
##                Df Deviance    AIC     LRT  Pr(>Chi)
## <none>              2393.3 2433.3
## perHoush        1   2496.8 2534.8 103.447 < 2.2e-16 ***
## IncMil          1   2518.9 2556.9 125.552 < 2.2e-16 ***
## State:pctWhite  8   2441.5 2465.5  48.146 9.264e-08 ***
```

```
## “Deviance=2393.345”
## “R2=0.1848755”
```

Modello con interazioni



Modello senza interazioni



Test Hosmer-Lemshow del modello senza interazioni

```
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  log4$y, fitted(log4)
## X-squared = 42.672, df = 8, p-value = 1.013e-06
```

Test Hosmer-Lemshow del modello con interazioni

```
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  log6$y, fitted(log6)
## X-squared = 19.8, df = 8, p-value = 0.01112
```