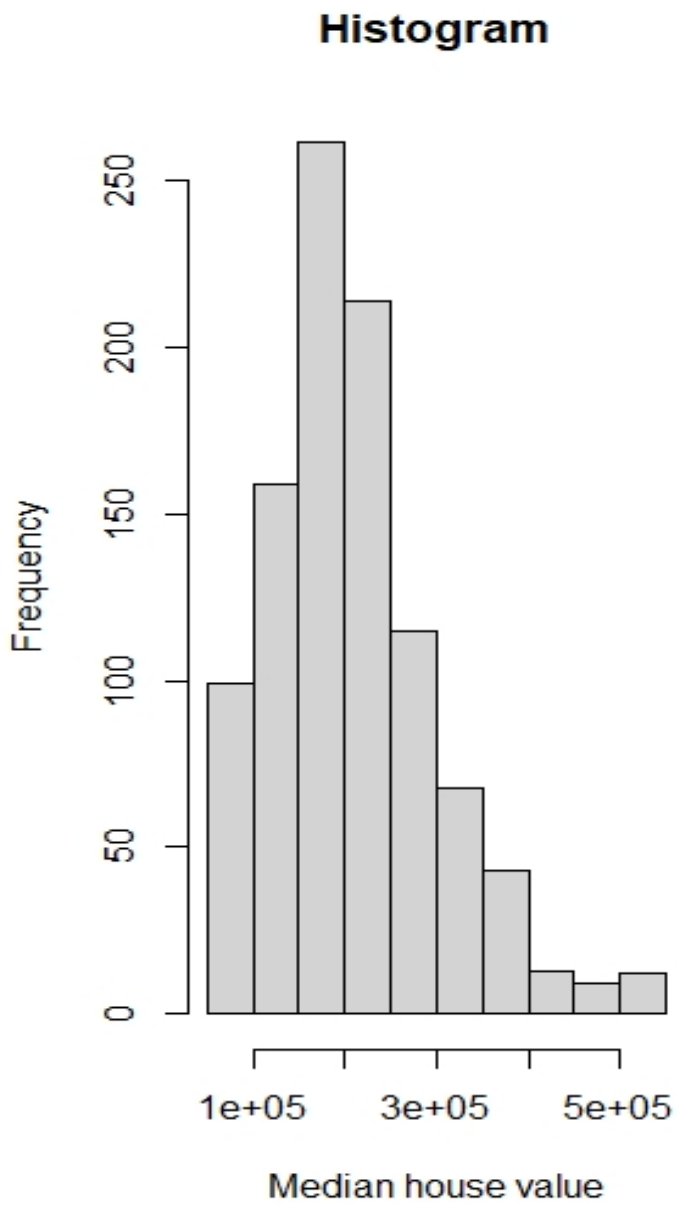
CALIFORNIA HOUSING PRICES

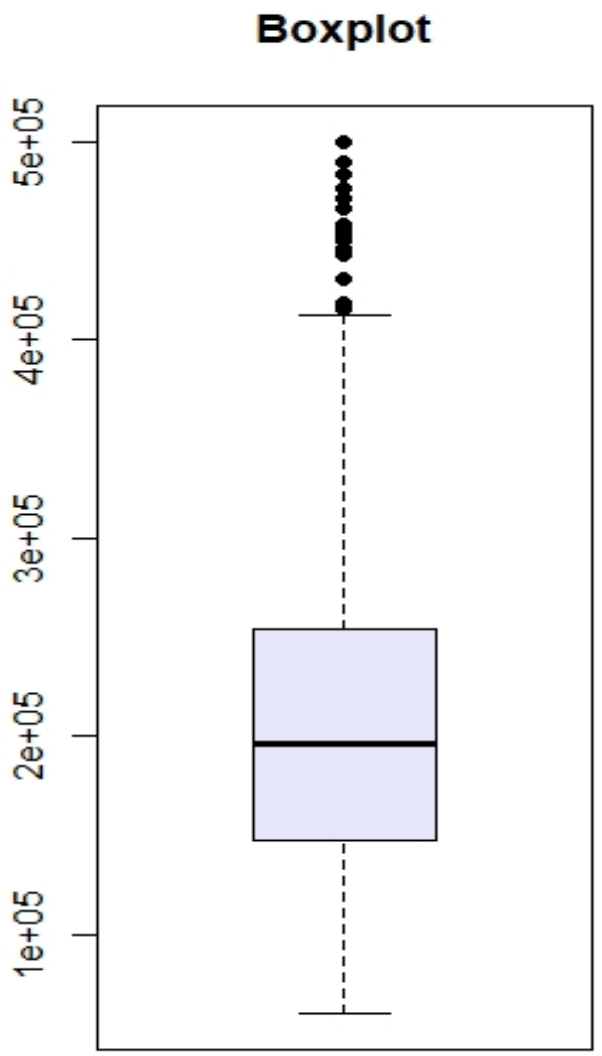
Corso di **Modelli e Metodi per l'Inferenza Statistica**
A.A. 2020/2021
Lorenzo Ferrara – Matteo Ghesini – Viviana Giorgi

https://www.kaggle.com/camnugent/california-housing-prices

1110 osservazioni (divise in test set e training set) sul valore medio delle case in California nel 1990: *Median_house_value* sarà la nostra variabile risposta

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
1	-122.23	37.88	41	880	129	322	126	8.3252	452600
2	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500
3	-122.24	37.85	52	1467	190	496	177	7.2574	352100
4	-122.25	37.85	52	1274	235	558	219	5.6431	341300
5	-122.25	37.85	52	1627	280	565	259	3.8462	342200
6	-122.25	37.85	52	919	213	413	193	4.0368	269700
7	-122.25	37.84	52	2535	489	1094	514	3.6591	299200
8	-122.25	37.84	52	3104	687	1157	647	3.1200	241400
9	-122.26	37.84	42	2555	665	1206	595	2.0804	226700
10	-122.25	37.84	52	3549	707	1551	714	3.6912	261100





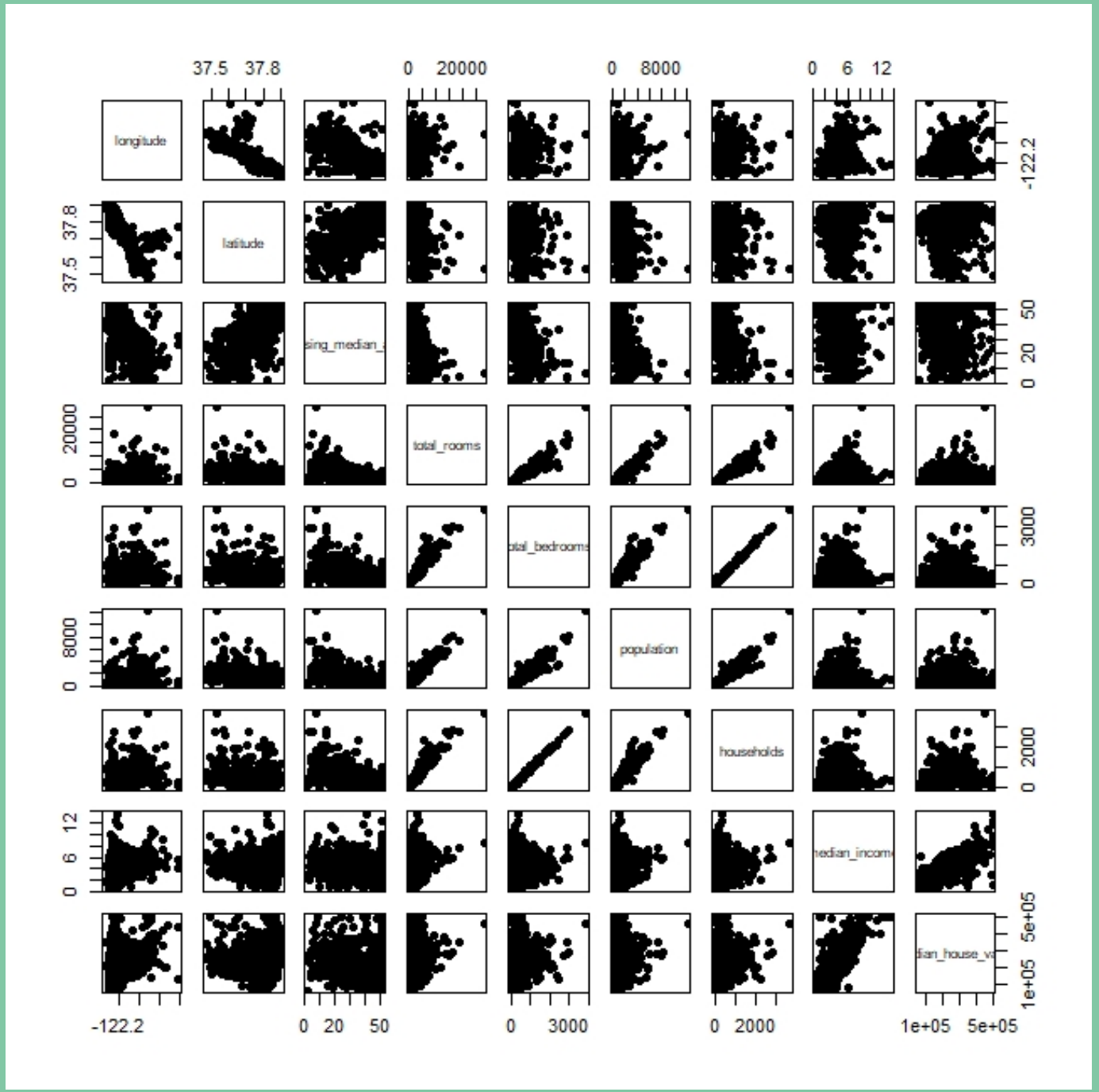
Statistica descrittiva

- Valore medio 200 mila dollari (da aggiustare con l'inflazione)
- Boxplot soddisfacente: pochi valori fuori dal 'baffo' destro ($q_3 + 1.5*(q_3-q_1)$)

Grafico con comando pairs

Le covariate (in ordine, dall'alto verso il basso):

Longitude
Latitude
Housing_median_age
Total_rooms
Total_bedrooms
Population
Households
Median_income



Primo modello contenente 8 covariate

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.358e+05  2.751e+06   0.158   0.8741
longitude           6.539e+04  2.786e+04   2.347   0.0191 *
latitude            2.013e+05  3.283e+04   6.131 1.26e-09 ***
housing_median_age  1.621e+02  1.963e+02   0.826   0.4090
total_rooms         2.752e+00  3.420e+00   0.805   0.4211
total_bedrooms     -3.897e+01  5.514e+01  -0.707   0.4799
population         -2.852e+01  6.386e+00  -4.467 8.87e-06 ***
households          1.316e+02  6.157e+01   2.137   0.0329 *
median_income       3.796e+04  1.286e+03  29.524  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53330 on 985 degrees of freedom
Multiple R-squared:  0.6452,    Adjusted R-squared:  0.6424
F-statistic: 223.9 on 8 and 985 DF,  p-value: < 2.2e-16
```

```
AIC(g) # 24469.87
```

Riduzione del modello tramite la **tecnica STEPWISE**

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.730e+05  2.457e+06  -0.193   0.8474
longitude      6.130e+04  2.669e+04   2.297   0.0218 *
latitude       2.122e+05  3.100e+04   6.845 1.34e-11 ***
population    -2.469e+01  5.205e+00  -4.743 2.42e-06 ***
households     9.337e+01  1.368e+01   6.826 1.52e-11 ***
median_income  3.866e+04  1.034e+03  37.397  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53310 on 988 degrees of freedom
Multiple R-squared:  0.6445,    Adjusted R-squared:  0.6427
F-statistic: 358.2 on 5 and 988 DF,  p-value: < 2.2e-16
```

5 covariate tutte significative
Diminuzione dell'AIC (smaller is better)

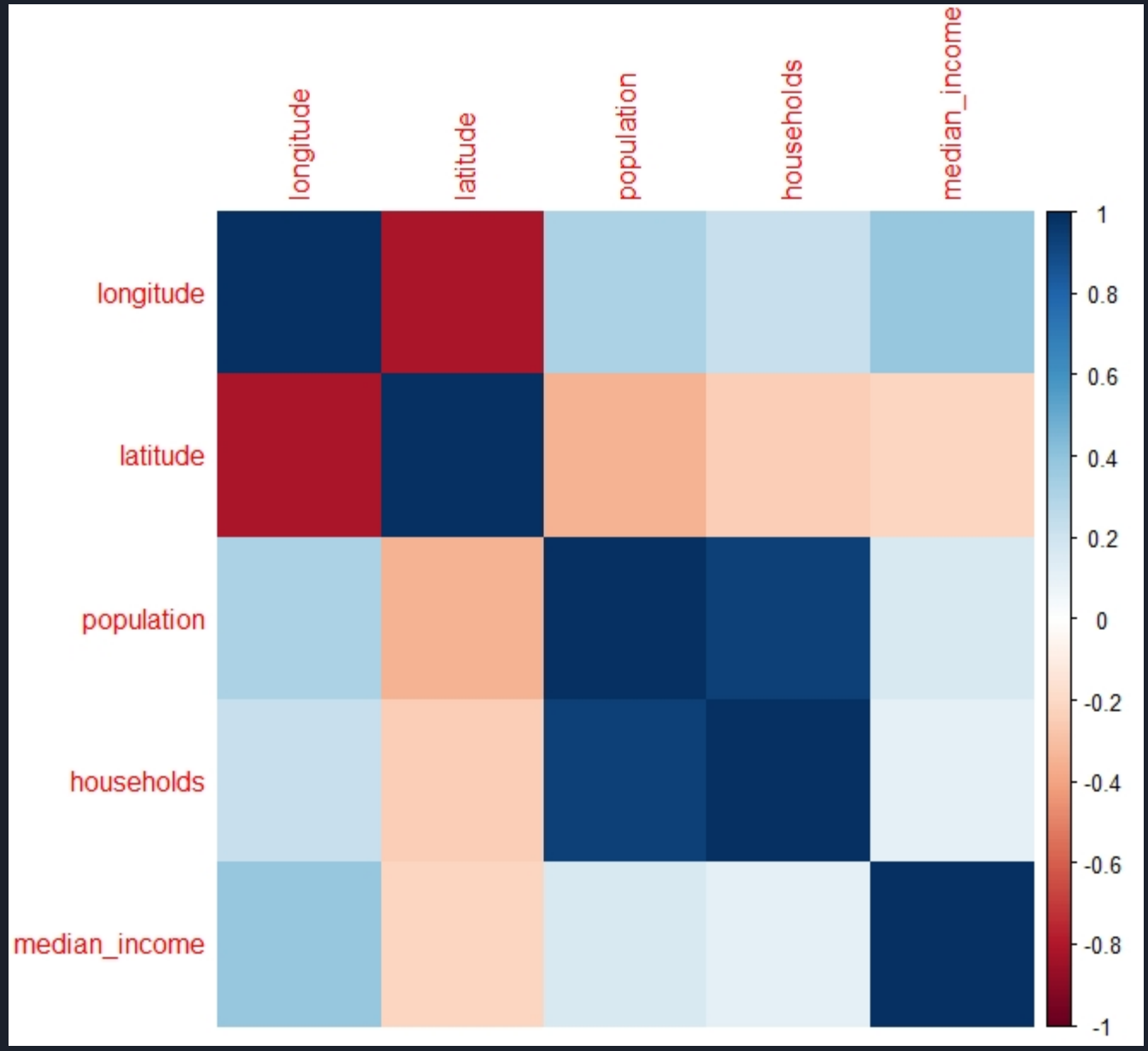
```
> AIC(g) # 24465.94
```

Studio della collinearità

Forte correlazione tra:

- latitude e longitude
- households e population

Aggiunta delle **interazioni** tra covariate



Modello con entrambe le interazioni

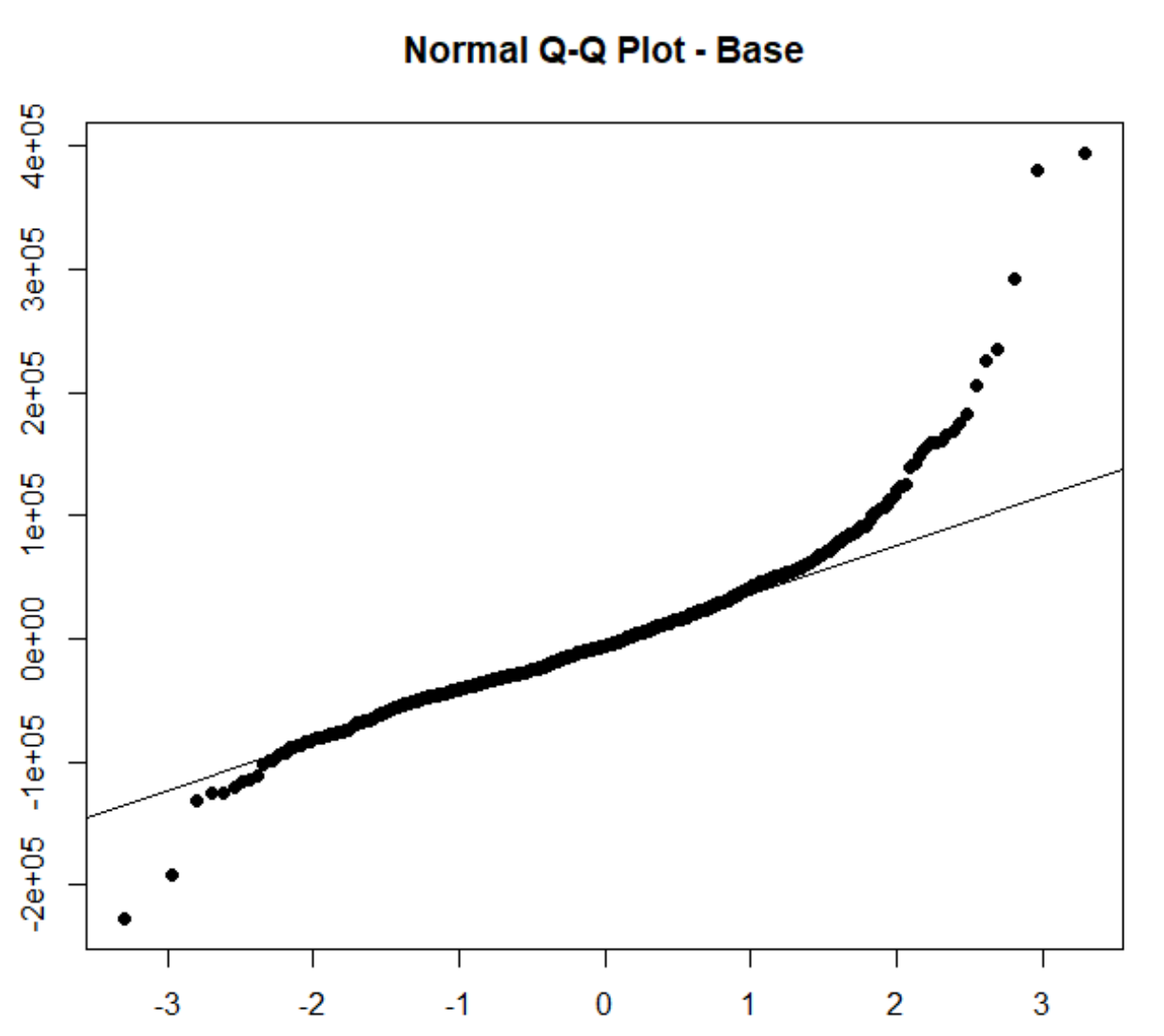
```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               7.143e+09  8.018e+08   8.908  < 2e-16 ***
longitude                 5.855e+07  6.565e+06   8.918  < 2e-16 ***
latitude                 -1.891e+08  2.125e+07  -8.899  < 2e-16 ***
population               -2.322e+01  5.788e+00  -4.011 6.49e-05 ***
households                8.954e+01  1.322e+01   6.774 2.15e-11 ***
median_income             3.762e+04  1.006e+03  37.397  < 2e-16 ***
longitude:latitude       -1.550e+06  1.740e+05  -8.909  < 2e-16 ***
population:households    -1.186e-03  1.532e-03  -0.774    0.439
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51310 on 986 degrees of freedom
Multiple R-squared:  0.6713,    Adjusted R-squared:  0.669
F-statistic: 287.7 on 7 and 986 DF,  p-value: < 2.2e-16
```

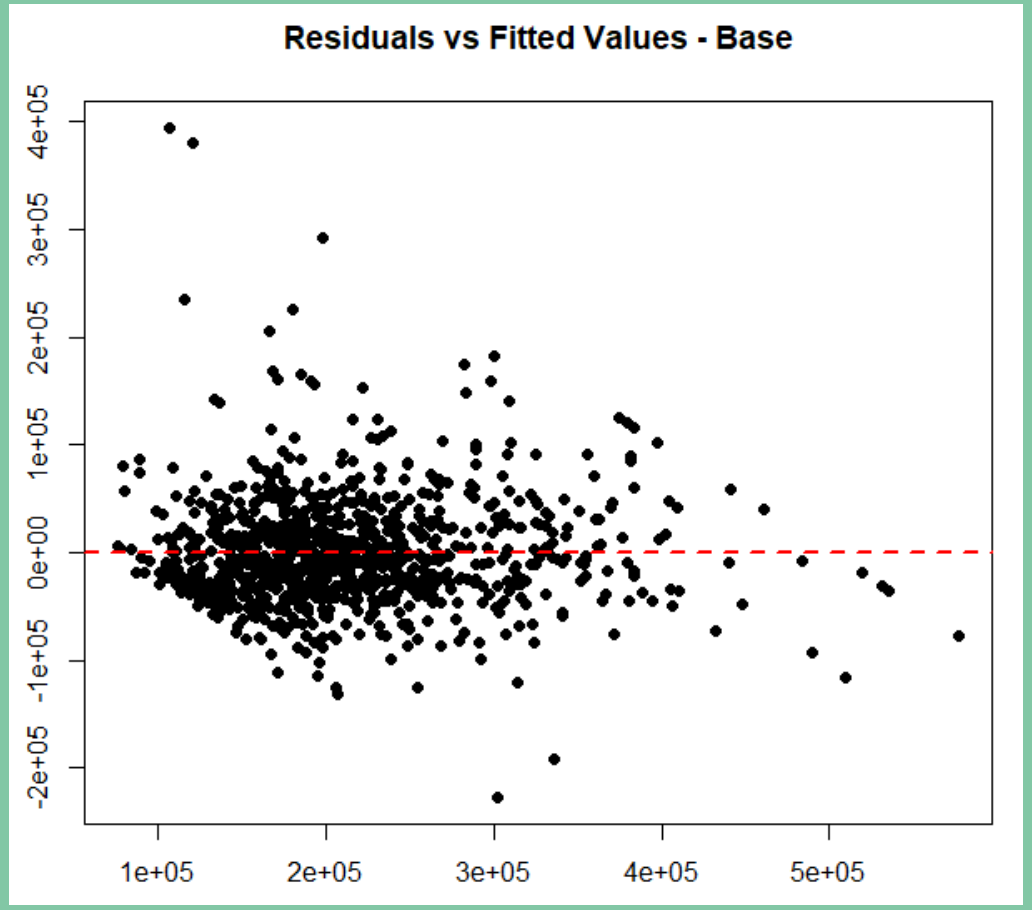
Modello finale con soltanto longitude*latitude

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.165e+09  8.012e+08   8.943  < 2e-16 ***
longitude          5.873e+07  6.560e+06   8.953  < 2e-16 ***
latitude          -1.897e+08  2.124e+07  -8.934  < 2e-16 ***
population        -2.546e+01  5.010e+00  -5.082 4.46e-07 ***
households         8.871e+01  1.317e+01   6.735 2.79e-11 ***
median_income      3.756e+04  1.002e+03  37.467  < 2e-16 ***
longitude:latitude -1.555e+06  1.739e+05  -8.944  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51300 on 987 degrees of freedom
Multiple R-squared:  0.6711,    Adjusted R-squared:  0.6691
F-statistic: 335.7 on 6 and 987 DF,  p-value: < 2.2e-16
```

Normalità e omoschedasticità

Verifichiamo le **ipotesi di validità** del modello





SHAPIRO - WILKS TEST del modello originale

p-value < 2.2×10^{-16}

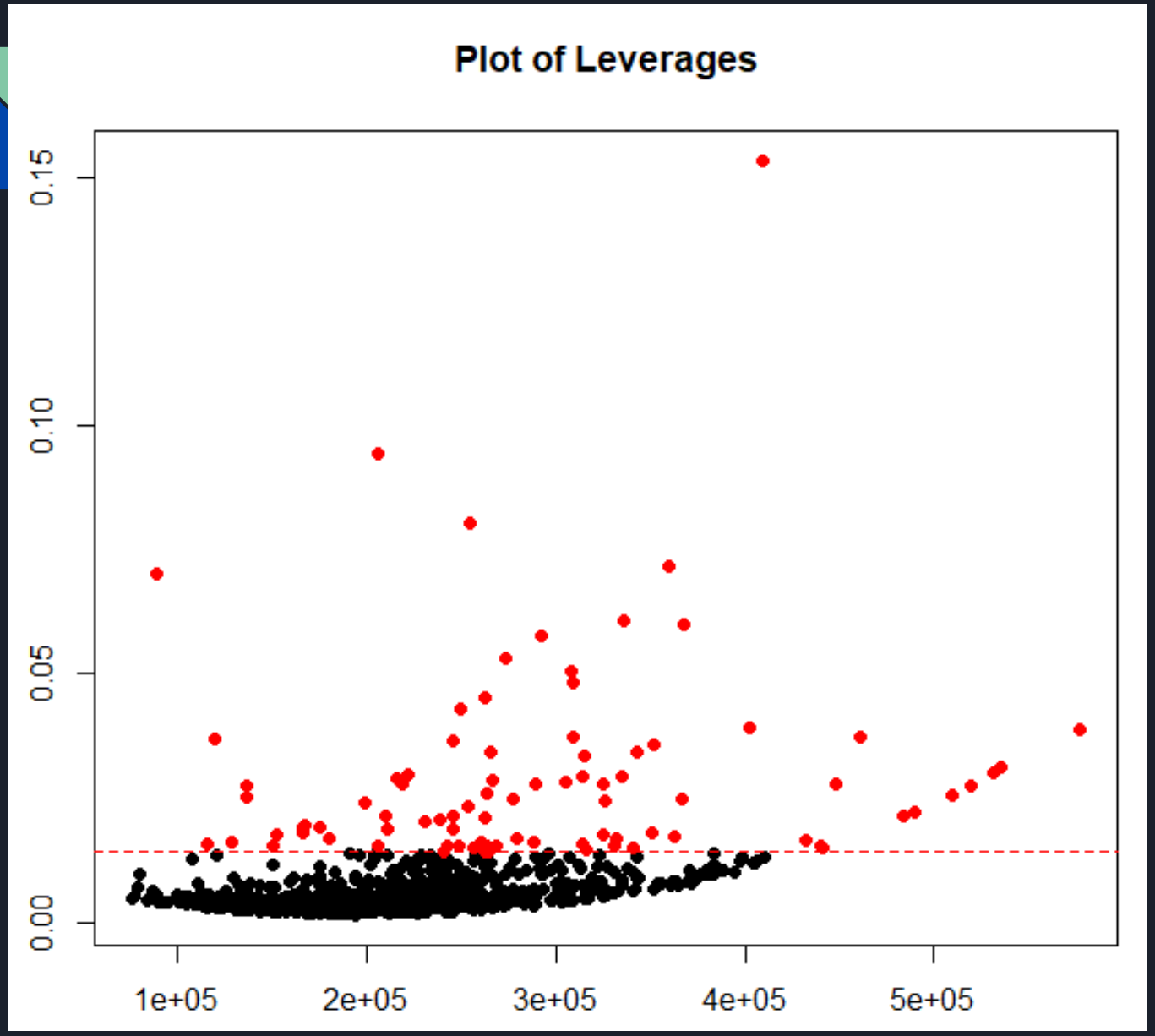
```
        Shapiro-Wilk normality test

data:  g$res
W = 0.90264, p-value < 2.2e-16
```

Goodness of fit

1. Leverages

1. Residui standardizzati

1. Distanza di Cook



1)LEVERAGES

Rimuovo i punti leva (87 osservazioni su 994)

Il nuovo modello risulta ugualmente soddisfacente

```
Residual standard error: 48370 on 900 degrees of freedom
Multiple R-squared:  0.6683,    Adjusted R-squared:  0.666
F-statistic: 302.2 on 6 and 900 DF,  p-value: < 2.2e-16
```

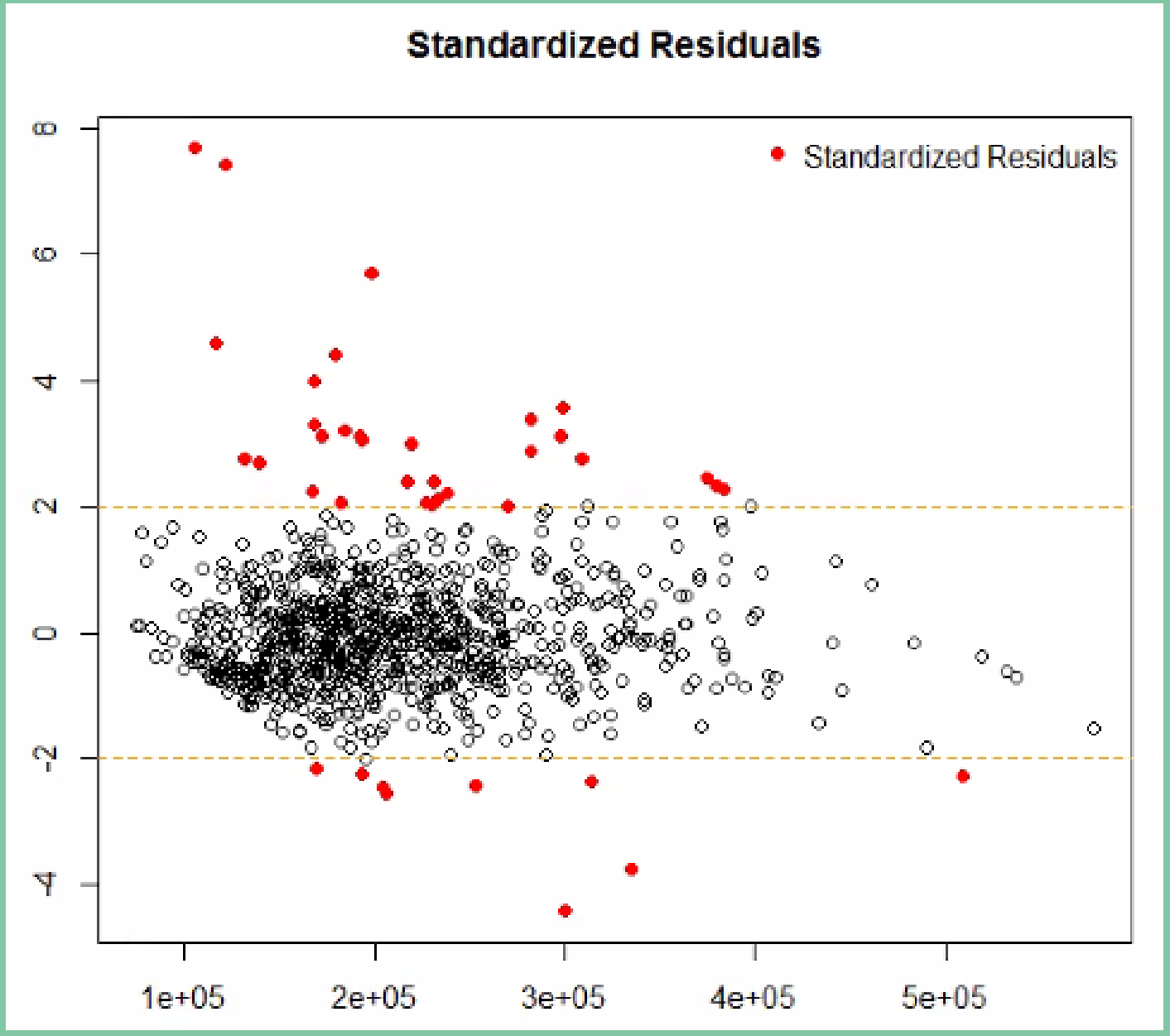
R^2 adj = 0.666

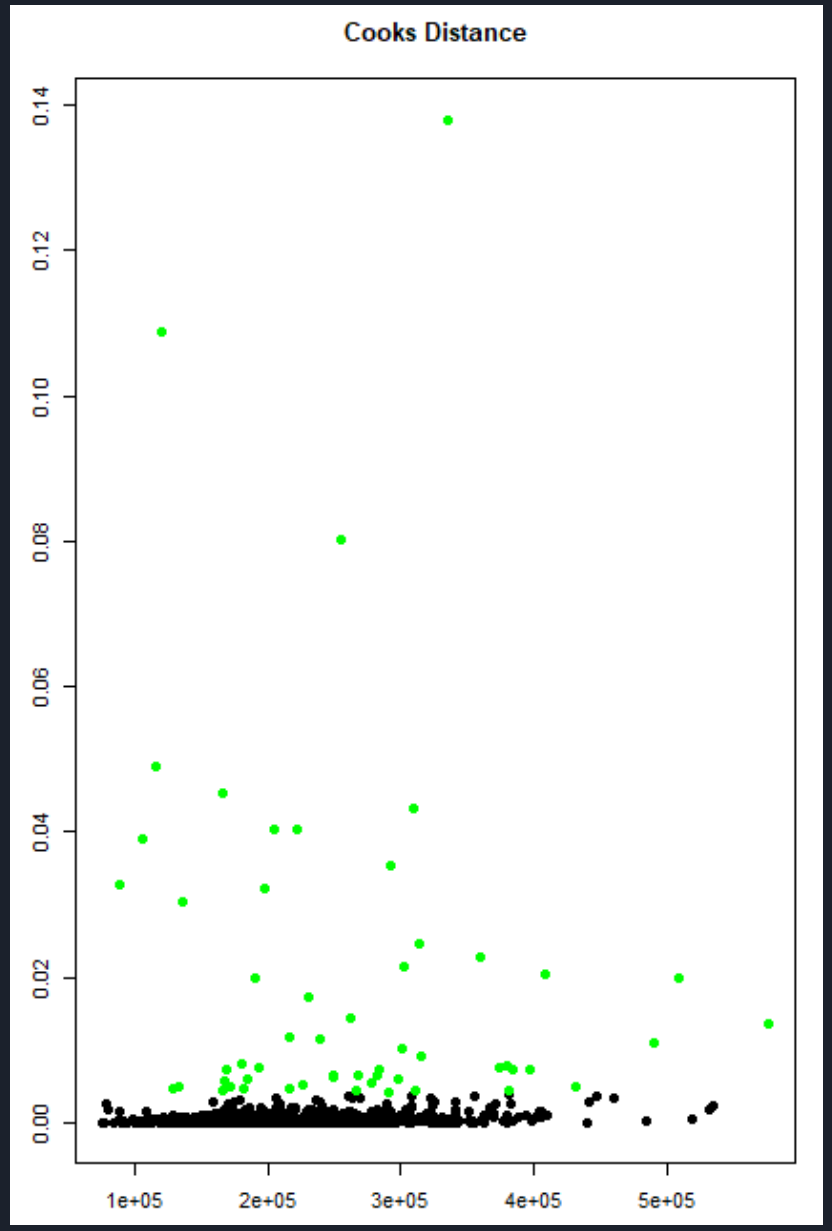
2) RESIDUI STANDARDIZZATI

Ripartiamo dal modello con 994 osservazioni

Rimuovo le 40 osservazioni in **rosso**
R^2 adj= 0.7969
Modello risulta migliorato!

```
Residual standard error: 37450 on 947 degrees of freedom
Multiple R-squared:  0.7981,    Adjusted R-squared:  0.7969
F-statistic: 624.1 on 6 and 947 DF,  p-value: < 2.2e-16
```





3) DISTANZA DI COOK

Rimuovo 55 osservazioni

Il modello risulta migliorato

R^2adj: 0.6691 → 0.7976

```
Residual standard error: 36630 on 932 degrees of freedom
Multiple R-squared:  0.7989,    Adjusted R-squared:  0.7976
F-statistic: 616.9 on 6 and 932 DF,  p-value: < 2.2e-16
```

Aspetto dei dati rimossi

	longitude	latitude	population	households	median_income	median_house_value
5	-122.25	37.85	565	259	3.8462	342200
60	-122.29	37.82	94	57	2.5625	60000
61	-122.29	37.83	554	187	3.3929	75700
62	-122.29	37.82	86	23	6.1183	75000
90	-122.27	37.80	396	85	1.2434	500001

Sono casi limite, nei quali notiamo:

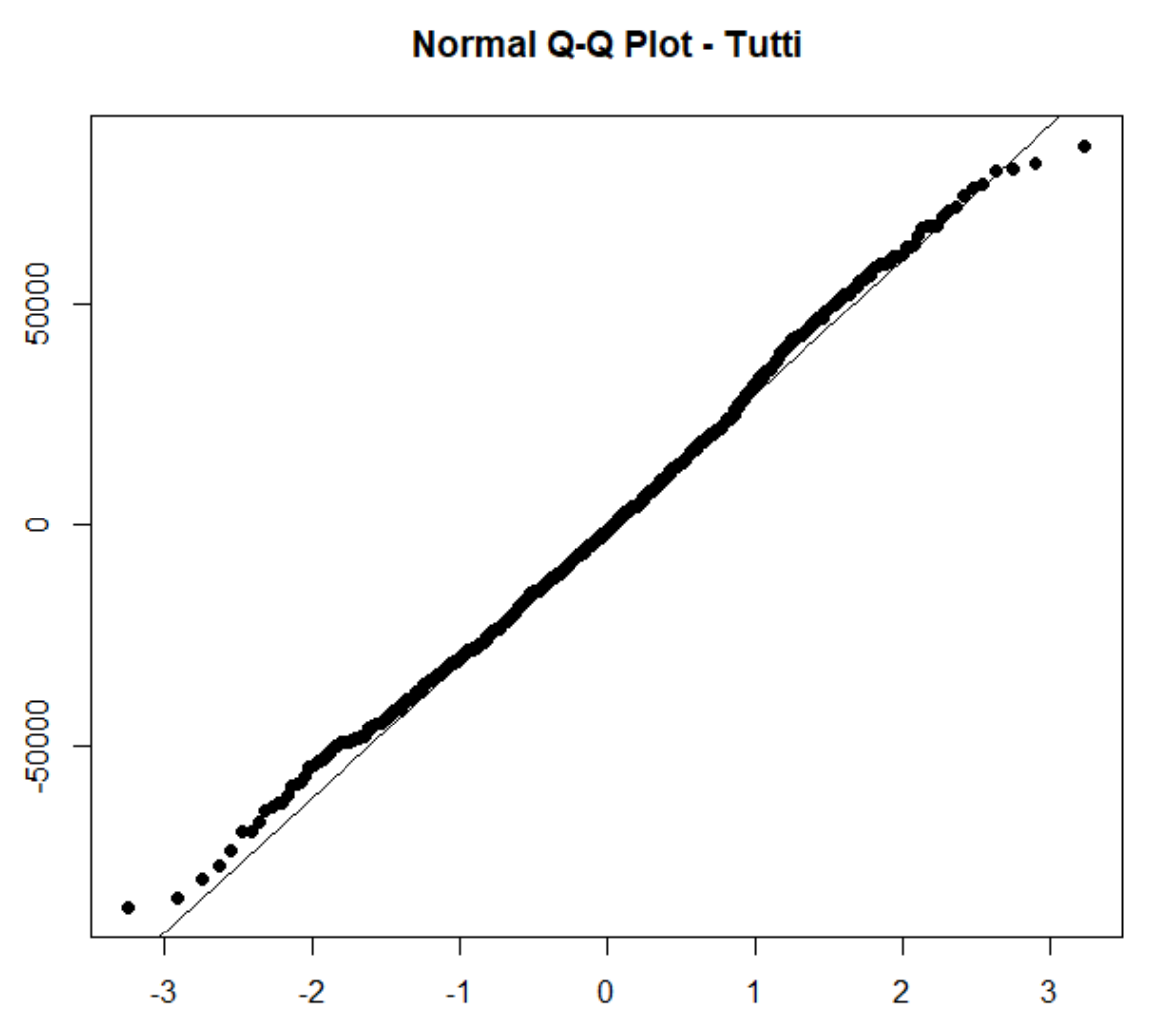
- grande differenza tra median_income e median_house_value
- zone tendenzialmente poco popolose (media = 1200)

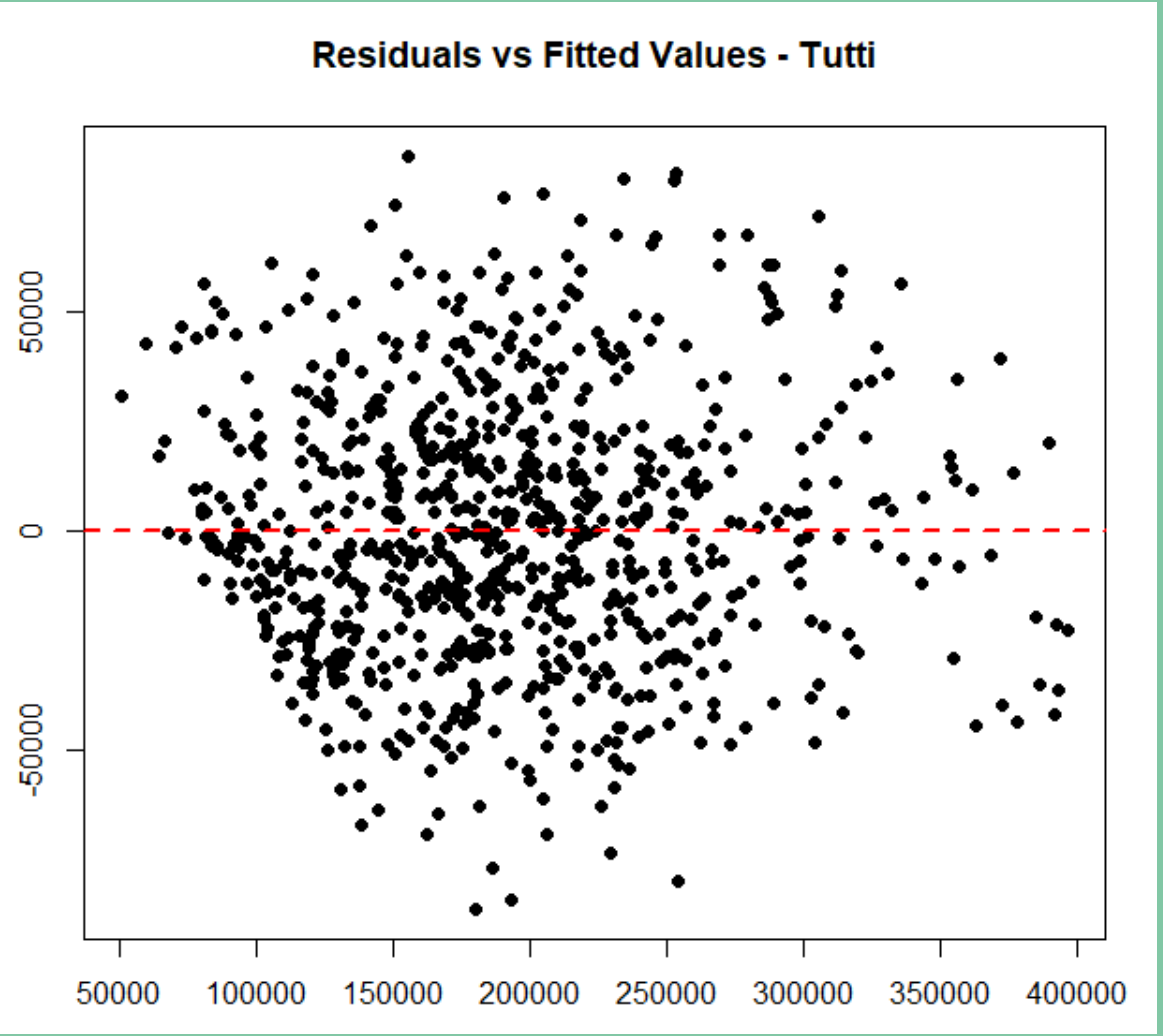
Combinando le tre tecniche

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.425e+09  6.072e+08   8.935   <2e-16 ***
longitude           4.444e+07  4.971e+06   8.940   <2e-16 ***
latitude           -1.438e+08  1.609e+07  -8.938   <2e-16 ***
population         -8.631e+01  5.163e+00 -16.716   <2e-16 ***
households          2.519e+02  1.267e+01  19.876   <2e-16 ***
median_income       4.035e+04  8.036e+02  50.207   <2e-16 ***
longitude:latitude -1.178e+06  1.317e+05  -8.942   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30170 on 814 degrees of freedom
Multiple R-squared:  0.8223,    Adjusted R-squared:  0.821
F-statistic: 627.7 on 6 and 814 DF,  p-value: < 2.2e-16
```

Il modello finale è molto soddisfacente: $R^2adj = 0.821$

Ho rimosso in totale 173 osservazioni → ne rimangono 821

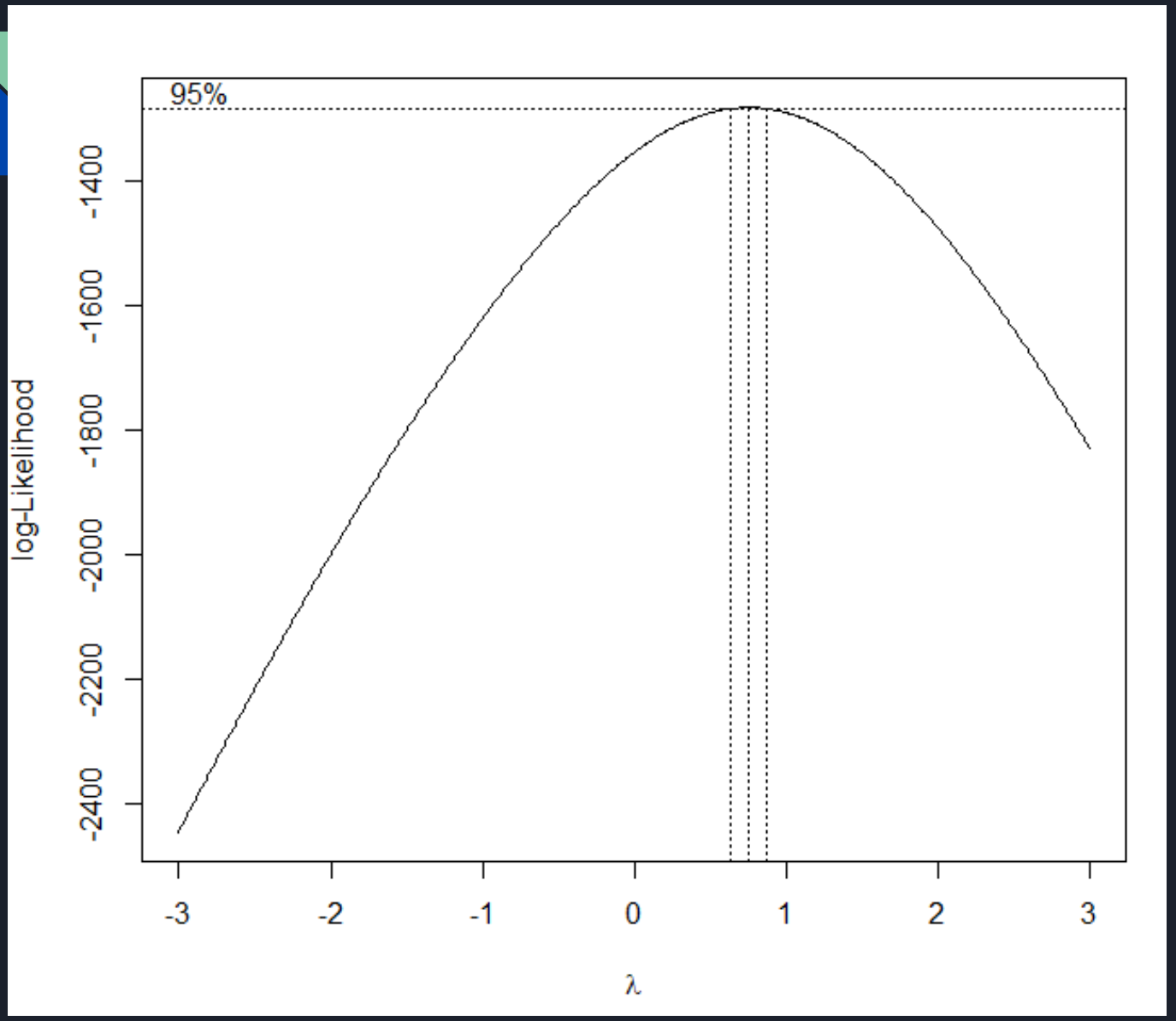




Aspetto del modello dopo aver rimosso i dati studiando Leverages, Residui Standardizzati, Distanza di Cook.

p-value è circa al 2.5% (molto migliorato ma non pienamente soddisfacente)

```
        Shapiro-Wilk normality test

data:  g_C_new$res
W = 0.9958, p-value = 0.02556
```



Trasformazione BOX-COX

R calcola il lambda che massimizza la verosimiglianza gaussiana dei residui.

$\lambda = 0.75$

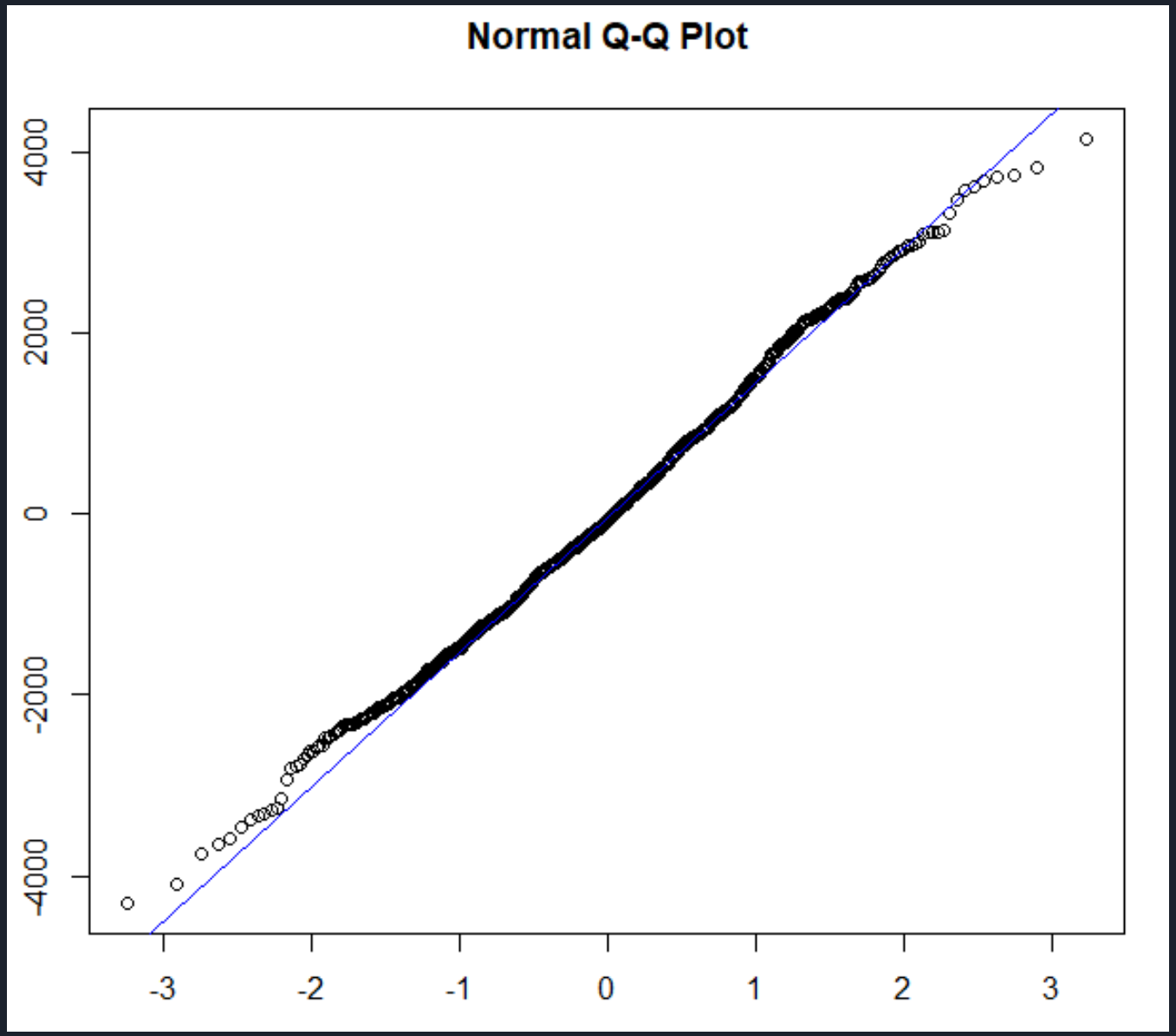
$Y^{new} = (Y^{\lambda} - 1) / \lambda$

dato che $\lambda \neq 0$

QQline del modello dopo BOX-COX

p-value sopra al 7%
→ soddisfacente

```
        Shapiro-Wilk normality test

data:  residuals(mod)
W = 0.99658, p-value = 0.07355
```



RIEPILOGO

$4/3*(\textbf{\textit{median\_house\_value}}^{3/4} - 1) = 2.62 * 10^8 + 2.146 * 10^6 * \textbf{\textit{longitude}} - 6.947 * 10^6 * \textbf{\textit{latitude}} - 5.689 * 10^4 * \textbf{\textit{longitude*latitude}} - 4.347 * \textbf{\textit{population}} + 12.66 * \textbf{\textit{households}} + 1.896 * 10^3 * \textbf{\textit{median\_income}}$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.620e+08  2.924e+07   8.962   <2e-16 ***
longitude         2.146e+06  2.394e+05   8.966   <2e-16 ***
latitude         -6.947e+06  7.748e+05  -8.965   <2e-16 ***
population       -4.347e+00  2.486e-01 -17.485   <2e-16 ***
households        1.266e+01  6.102e-01  20.754   <2e-16 ***
median_income     1.896e+03  3.869e+01  49.008   <2e-16 ***
longitude:latitude -5.689e+04  6.343e+03  -8.969   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1453 on 814 degrees of freedom
Multiple R-squared:  0.8177,    Adjusted R-squared:  0.8164
F-statistic: 608.6 on 6 and 814 DF,  p-value: < 2.2e-16
```

$R^2adj = 0.8164$

INTERVALLO DI PREVISIONE di livello 95%

Verifica della bontà tramite **cross-validazione** con 110 nuovi dati

Dati evidenziati sono esterni all'intervallo iniziale

FIT	LOWER	UPPER		FIT	LOWER	UPPER
28333	1771	72916		146831	93651	205402
35052	2191	81042		154202	100291	213362
41995	6253	89256		161499	106882	221237
49117	11070	97540		168715	113416	229021
56383	16389	105876		175843	119882	236707
63762	22072	114249		182877	126271	244291
71229	28033	122644		189810	132574	251768
78762	34207	131050		196636	138783	259134
86340	40546	139452		203350	144891	266384
93947	47012	147842		209946	150890	273514
101566	53573	156207		216419	156774	280522
109184	60201	164539		222766	162536	287403
116787	66875	172829		228980	168170	294154
124364	73574	181067		235059	173671	300772
131903	80279	189247		240997	179032	307255
139396	86977	197361		246792	184248	313600

Errore quadratico medio di previsione

```
> MSPE = mean( (dati_C_new[ seq(1,110,1),6] - fit)^2 )
> MSPE
[1] 12077309836
> sqrt(MSPE)
[1] 109896.8
```

La deviazione standard è circa 100mila a fronte di una variabilità iniziale tra 60mila e 500mila

Il modello risulta accettabile