## Diagnostyka modelu

$$\begin{aligned}\hat{\varepsilon} = y - \hat{y} &= (I-H)y \\ &= (I-H)X\beta + (I-H)\varepsilon \\ &= (I-H)\varepsilon\end{aligned}$$

$$\text{var}\,\hat{\varepsilon} = \text{var}\,(I-H)\varepsilon = (I-H)\sigma^2$$

$$\text{var}\,\hat{\varepsilon}_i = \sigma^2(1-h_i)$$

Dowód [5.4]

$$\sum_i h_i = p \qquad h_i \geq 1/n \quad \forall i$$

Dowód [5.5-5.6]

Przykład

```
> head(savings)
             sr pop15 pop75     dpi ddpi
Australia 11.43 29.35  2.87 2329.68 2.87
Austria   12.07 23.32  4.41 1507.99 3.93
Belgium   13.17 23.80  4.43 2108.47 3.82
Bolivia    5.75 41.89  1.67  189.13 0.22
Brazil    12.88 42.19  0.83  728.47 4.56
Canada     8.79 31.72  2.85 2982.88 2.43
…
```

sr - savings rate - personal saving divided by disposable income

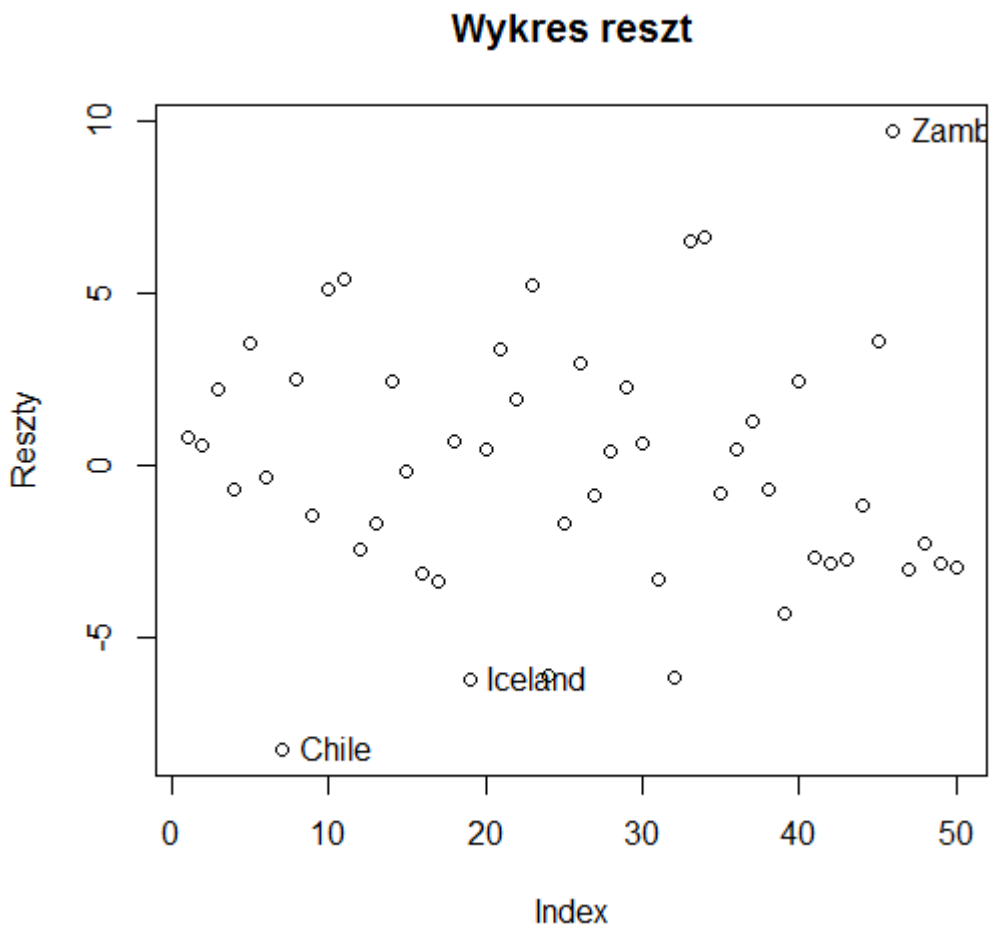
pop15 - percent population under age of 15

pop75 - percent population over age of 75

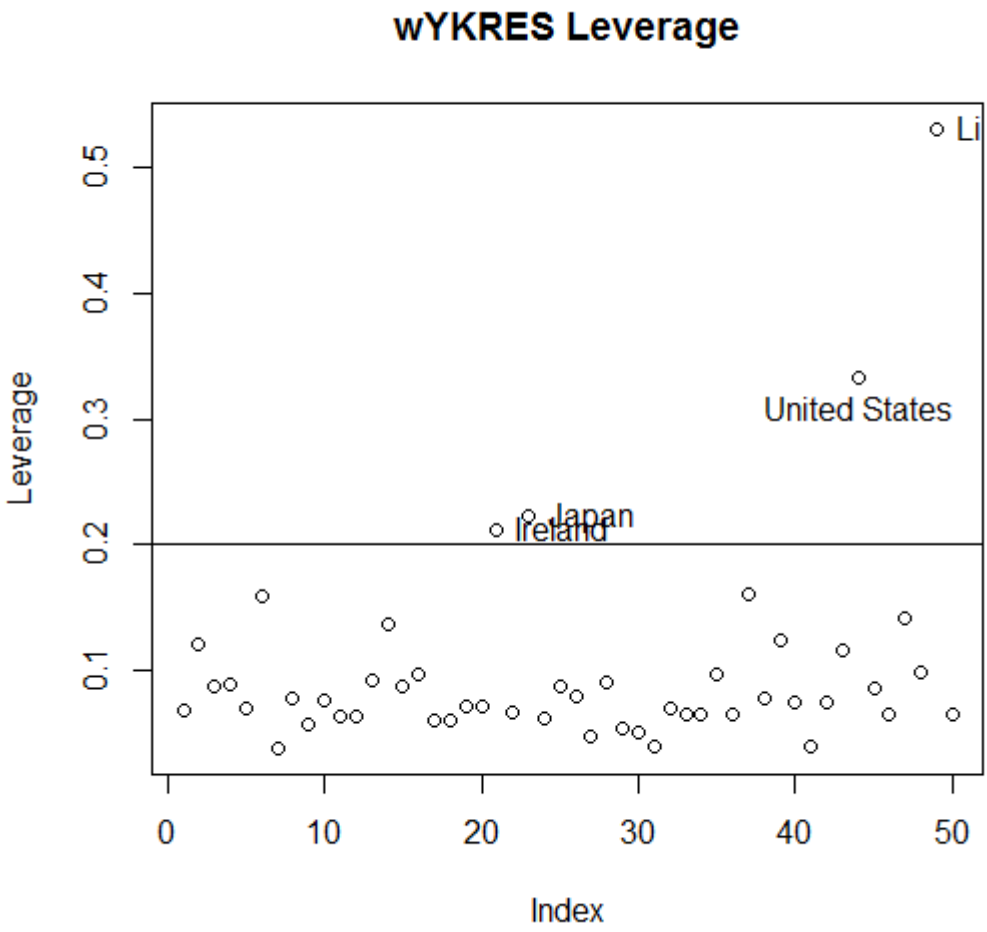
dpi - per-capita disposable income in dollars

ddpi - percent growth rate of dpi

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> plot(g$res,ylab="Reszty",main="Wykres reszt")
```




```
> x <- model.matrix(g)
> (lev <- hat(x))
 [1] 0.06771343 0.12038393 0.08748248 0.08947114 0.06955944 0.15840239
0.03729796
 [8] 0.07795899 0.05730171 0.07546780 0.06271782 0.06372651 0.09204246
0.13620478
[15] 0.08735739 0.09662073 0.06049212 0.06008079 0.07049590 0.07145213
0.21223634
[22] 0.06651170 0.22330989 0.06079915 0.08634787 0.07940290 0.04793213
0.09061400
[29] 0.05421789 0.05035056 0.03897459 0.06937188 0.06504891 0.06425415
0.09714946
[36] 0.06510405 0.16080923 0.07732854 0.12398898 0.07359423 0.03964224
0.07456729
[43] 0.11651375 0.33368800 0.08628365 0.06433163 0.14076016 0.09794717
0.53145676
[50] 0.06523300

> plot(lev,ylab="Leverage",main="WYKRES Leverage")
> abline(h=2*5/50)
> names(lev) <- countries
> lev[lev > 0.2]
      Ireland         Japan United States         Libya
    0.2122363     0.2233099     0.3336880     0.5314568
```




```
> summary(g)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

```
> g.49.44 <- lm(sr ~ pop15 + pop75 + dpi + ddpi,
savings[setdiff(1:50,c(44,49)),])
> summary(g.49.44)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings[setdiff(1:50,
    c(44, 49)), ])

Residuals:
    Min      1Q  Median      3Q     Max
-8.0875 -2.5781 -0.1206  2.1944  9.3635

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.400e+01  8.433e+00   2.846  0.00675 **
pop15       -3.806e-01  1.623e-01  -2.345  0.02371 *
pop75       -1.322e+00  1.162e+00  -1.137  0.26175
dpi         -9.787e-05  1.123e-03  -0.087  0.93098
ddpi         6.170e-01  2.721e-01   2.267  0.02845 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.833 on 43 degrees of freedom
Multiple R-squared:  0.3543,    Adjusted R-squared:  0.2942
F-statistic: 5.899 on 4 and 43 DF,  p-value: 0.0007095
```

## Reszty studentyzowane

$$\operatorname{var} \hat{\varepsilon}_i = \sigma^2(1 - h_i)$$

Reszta studentyzowana:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Gdy model regresji jest poprawny, to

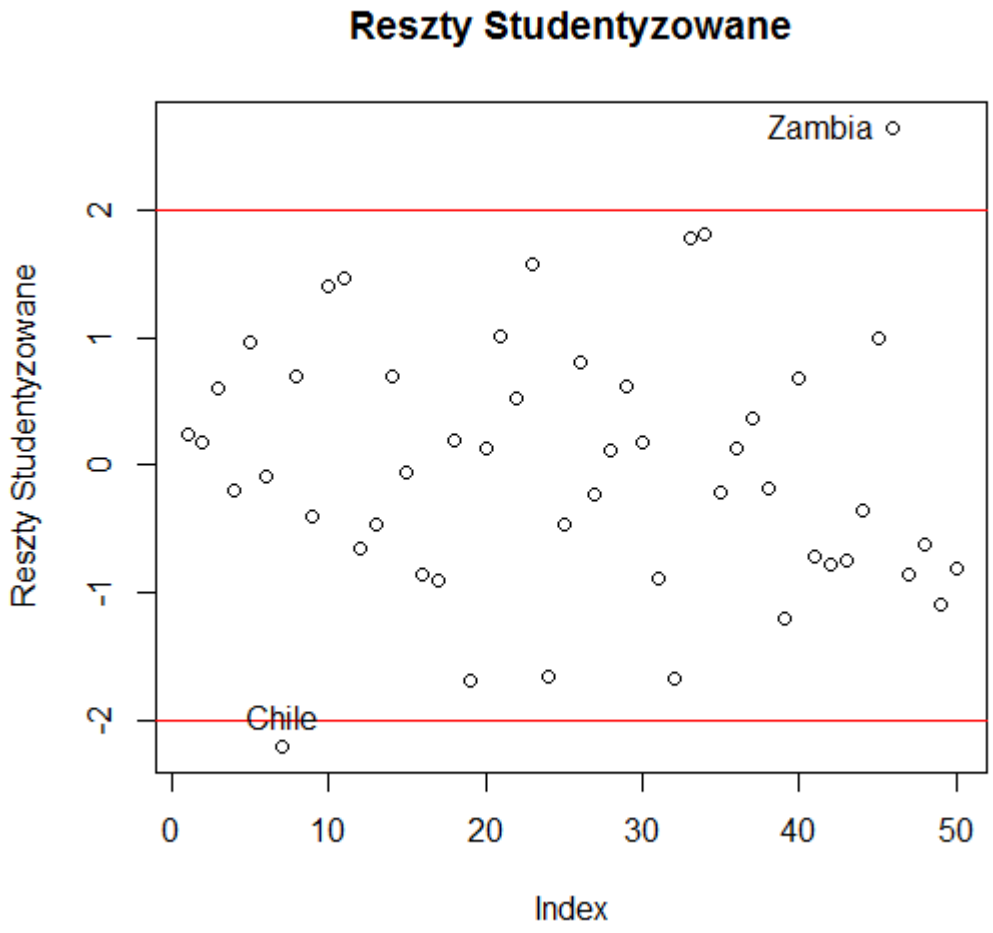
$$\operatorname{var} r_i = 1$$

oraz $r_i$ jest nieskorelowana z $r_j$.

Reszty studentyzowane wskazują na dane odległe od równania regresji (nieopisywane przez regresję). Przyjmuje się jako dużą wartość $|r_i|>2$

```
> (gs <- summary(g))

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
```

```
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904

> gs$sig
[1] 3.802669
> stud <- g$res/(gs$sig*sqrt(1-lev))
> names(stud) <- countries
> stud[abs(stud)>2]
    Chile    Zambia
-2.209074  2.650915
> cbind(savings$sr,g$fitted.values,g$residuals)[c(46,7),]
        [,1]     [,2]      [,3]
Zambia 18.56 8.809086  9.750914
Chile   0.60 8.842231 -8.242231
```

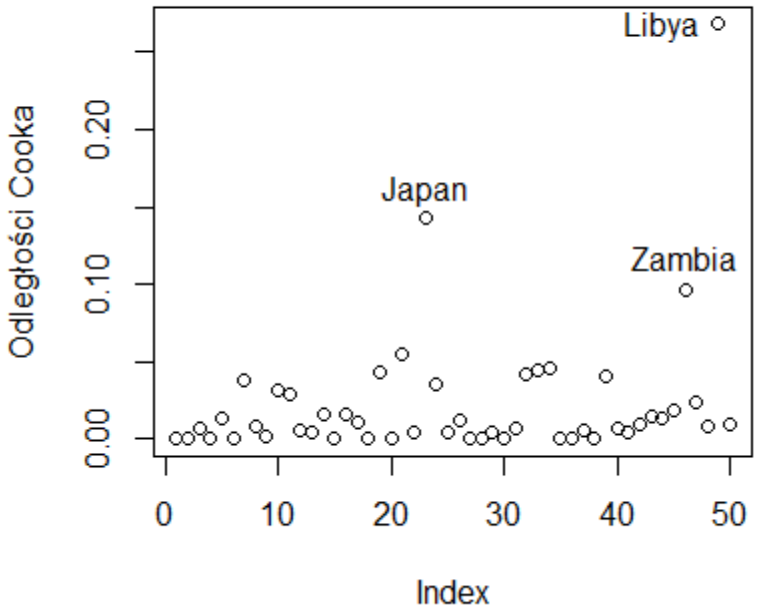



## Odległość Cooka

Prawdziwy wpływ i-tej obserwacji można zmierzyć efektem jej usunięcia.

$$D_i = \frac{\left\| \hat{y} - \widehat{y_{(i)}} \right\|^2}{p\widehat{\sigma^2}}$$

```
> cook <- cooks.distance(g)
> plot(cook,ylab="Odległości Cooka")
> identify(1:50,cook,countries)
[1] 23 46 49
```




Można pokazać, że

$$D_i = \frac{r_i^2 h_i}{p\,(1 - h_i)}$$

Po przekształceniu:

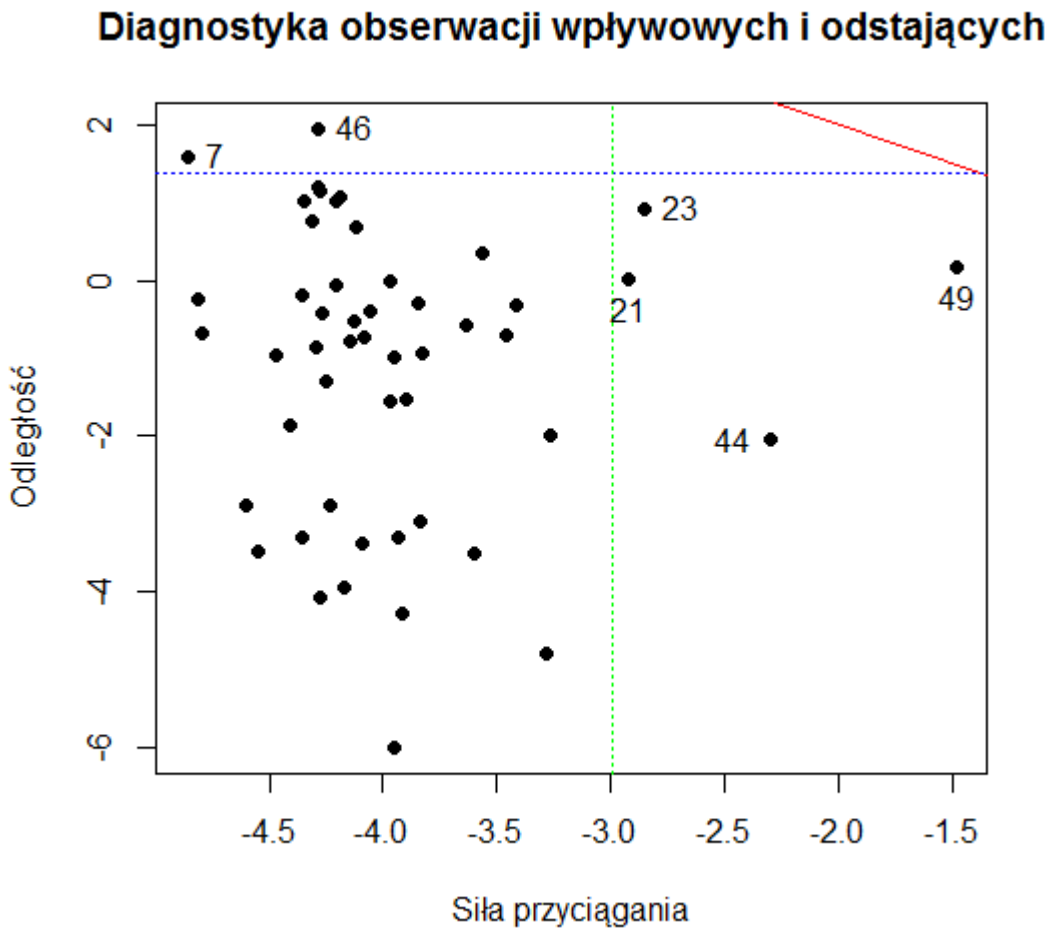
$$\log\left(r_i^2\right) = -\log\left(\frac{h_i}{p\,(1 - h_i)}\right) + \log\left(D_i\right)$$

Wartości nietypowe:

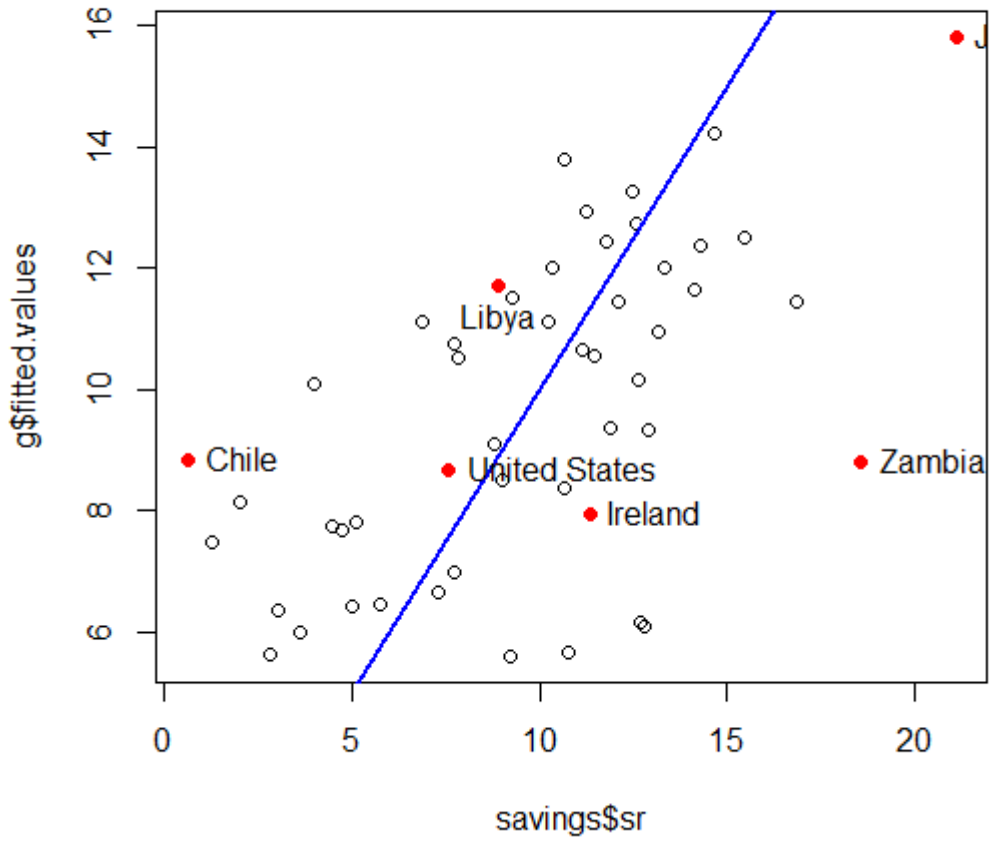
$r_i^2 > 4$ gdy punkt jest nieobjaśniony dobrze przez regresję

$h_i > \frac{2p}{n} \Longrightarrow \frac{h_i}{p\,(1 - h_i)} > \frac{2}{n - 2p}$ gdy punkt jest punktem wpływu

$D_i > 1$ gdy punkt modyfikuje prognozy




```
> wykresDiagnostyczny(g)
[1] Numery wyróżnionych punktów:
[1]  7 21 23 44 46 49
[1] Siła przyciągania:
       Chile      Ireland        Japan United States       Zambia        Libya
   -4.860243    -2.920936    -2.855919    -2.300989    -4.286648    -1.483444
[1] Odległość:
       Chile      Ireland        Japan United States       Zambia        Libya
  1.585147171  0.009477739  0.909722467 -2.053825827  1.949809984  0.166938873
                [,1]     [,2]
Chile           0.60  8.842231
Ireland        11.34  7.948869
Japan          21.10 15.818514
United States   7.56  8.671590
Zambia         18.56  8.809086
Libya           8.89 11.719526
```

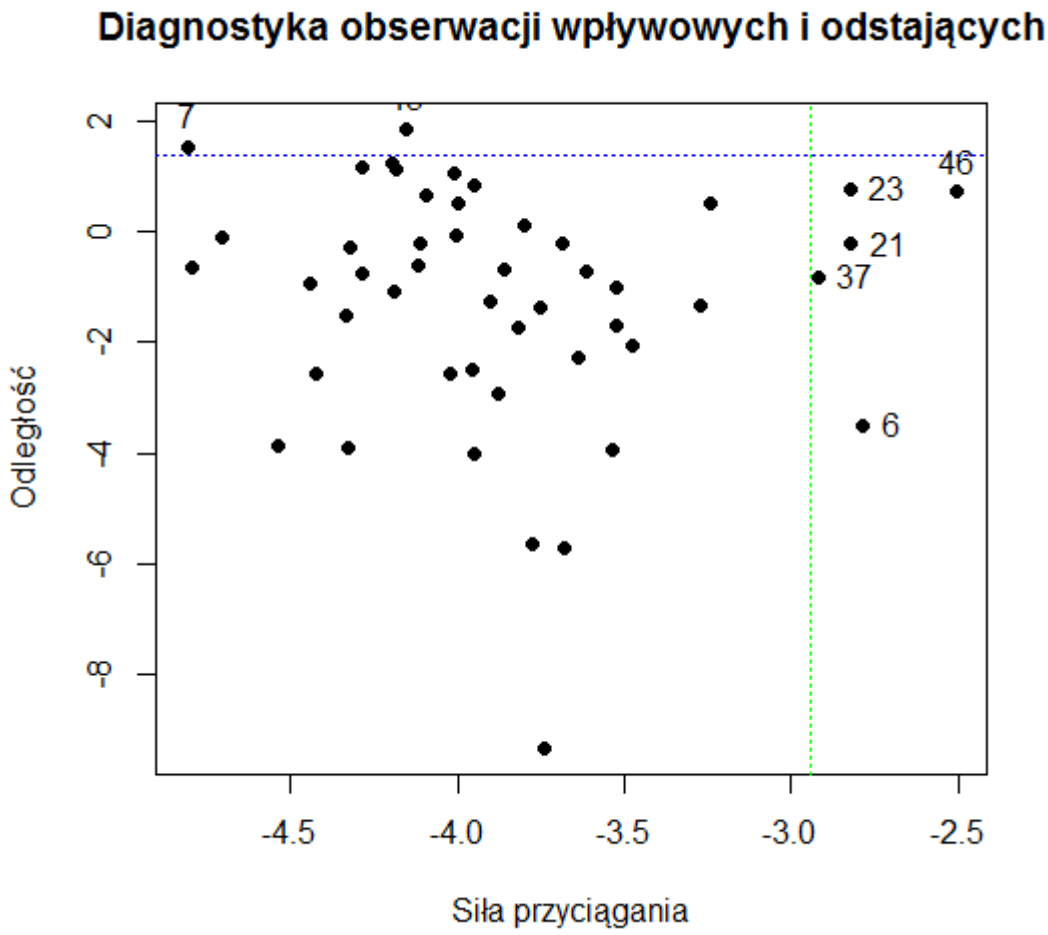



Po usunięciu 44 i 49 (USA i Libia)

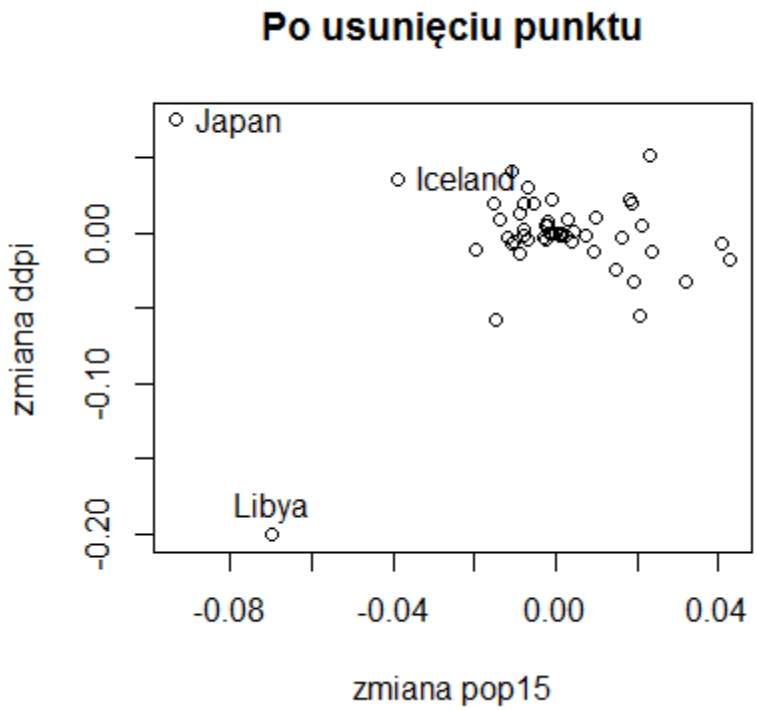
```
> summary(g)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904

> nowe <- setdiff(1:50,c(44,49))

> g.red <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings[nowe,])
> summary(g.red)
```

```
Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings[nowe,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-8.0875 -2.5781 -0.1206  2.1944  9.3635

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.400e+01  8.433e+00   2.846  0.00675 **
pop15       -3.806e-01  1.623e-01  -2.345  0.02371 *
pop75       -1.322e+00  1.162e+00  -1.137  0.26175
dpi         -9.787e-05  1.123e-03  -0.087  0.93098
ddpi         6.170e-01  2.721e-01   2.267  0.02845 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.833 on 43 degrees of freedom
Multiple R-squared:  0.3543,    Adjusted R-squared:  0.2942
F-statistic: 5.899 on 4 and 43 DF,  p-value: 0.0007095
```

| | Estimate | | Pr(>\|t\|) | |
|---|---|---|---|---|
| (Intercept) | 28.5661 | 24.0000 | 0.000334 *** | 0.00675 ** |
| pop15 | -0.4612 | -0.3806 | 0.002603 ** | 0.02371 * |
| pop75 | -1.6915 | -1.3220 | 0.125530 | 0.26175 |
| dpi | -0.0003 | -0.0001 | 0.719173 | 0.93098 |
| ddpi | 0.4097 | 0.6170 | 0.042471 * | 0.02845 * |

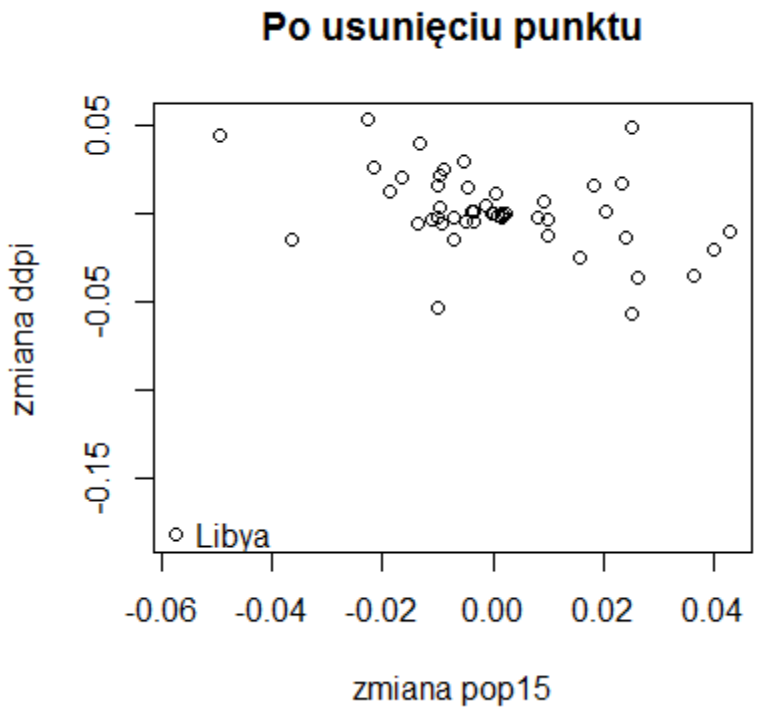



```
> wykresDiagnostyczny(g.red)
[1] Numery wyróżnionych punktów:
[1]  6  7 21 23 37 45 46
[1] Siła przyciągania:
        Canada          Chile        Ireland          Japan South Rhodesia
     -2.784520      -4.808544      -2.823132      -2.824710      -2.920998
        Zambia        Jamaica
     -4.154912      -2.506116
[1] Odległość:
        Canada          Chile        Ireland          Japan South Rhodesia
    -3.5105437      1.5333444    -0.2278992      0.7754542     -0.8296246
       Zambia         Jamaica
     1.8618759      0.7151247


> ginf <- lm.influence(g)
```




```
> gj <- lm(sr~pop15+pop75+dpi+ddpi,savings,subset=(countries != "Japan"))
> summary(gj)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,
    subset = (countries != "Japan"))

Residuals:
   Min     1Q Median     3Q    Max
-7.997 -2.592 -0.115  2.032 10.157

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.9401714  7.7839968   3.076  0.00361 **
pop15       -0.3679015  0.1536296  -2.395  0.02096 *
pop75       -0.9736743  1.1554502  -0.843  0.40397
dpi         -0.0004706  0.0009191  -0.512  0.61116
ddpi         0.3347486  0.1984457   1.687  0.09871 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.738 on 44 degrees of freedom
Multiple R-squared:  0.277,     Adjusted R-squared:  0.2113
F-statistic: 4.214 on 4 and 44 DF,  p-value: 0.005649
```

| | Estimate | | Pr(>\|t\|) | |
|---|---|---|---|---|
| (Intercept) | 28.5661 | 23.9402 | 0.000334 *** | 0.00361 ** |
| pop15 | -0.4612 | -0.3692 | 0.002603 ** | 0.02096 * |
| pop75 | -1.6915 | --0.9737 | 0.125530 | 0.40397 |
| dpi | -0.0003 | -0.0005 | 0.719173 | 0.61116 |
| ddpi | 0.4097 | 0.3347 | 0.042471 * | 0.09871 . |




```
> gjl <- lm(sr~pop15+pop75+dpi+ddpi,savings[setdiff(1:50,c(23,49)),])
> summary(gjl)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings[setdiff(1:50,
    c(23, 49)), ])

Residuals:
    Min      1Q  Median      3Q     Max
-7.8560 -2.8171  0.1004  2.1668  9.8014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.5802042  8.5112236   2.418   0.0199 *
pop15       -0.3105884  0.1644866  -1.888   0.0658 .
pop75       -0.6446624  1.2039200  -0.535   0.5951
dpi         -0.0004480  0.0009198  -0.487   0.6287
ddpi         0.5166284  0.2719756   1.900   0.0642 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.74 on 43 degrees of freedom
Multiple R-squared:  0.2925,    Adjusted R-squared:  0.2267
F-statistic: 4.444 on 4 and 43 DF,  p-value: 0.004285
```

| | Estimate | | | Pr(>\|t\|) | | |
|---|---|---|---|---|---|---|
| (Intercept) | 28.5661 | 23.9402 | 20.5802 | 0.0003 *** | 0.004 ** | 0.02 * |
| pop15 | -0.4612 | -0.3692 | -0.3106 | 0.003 ** | 0.02 * | 0.07 . |
| pop75 | -1.6915 | --0.9737 | -0.6447 | 0.13 | 0.40 | 0.60 |
| dpi | -0.0003 | -0.0005 | -0.0004 | 0.72 | 0.61 | 0.63 |
| ddpi | 0.4097 | 0.3347 | 0.5166 | 0.04 * | 0.10 . | 0.06 . |