Prezentacja i wizualizacja danych

Organizacyjnie

Prowadzący:

dr Mariusz Rafało

mrafalo@sgh.waw.pl

http://mariuszrafalo.pl (hasło:WIZ)

Regresja liniowa

Regresja liniowa

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, i = 1, 2, ..., n$$

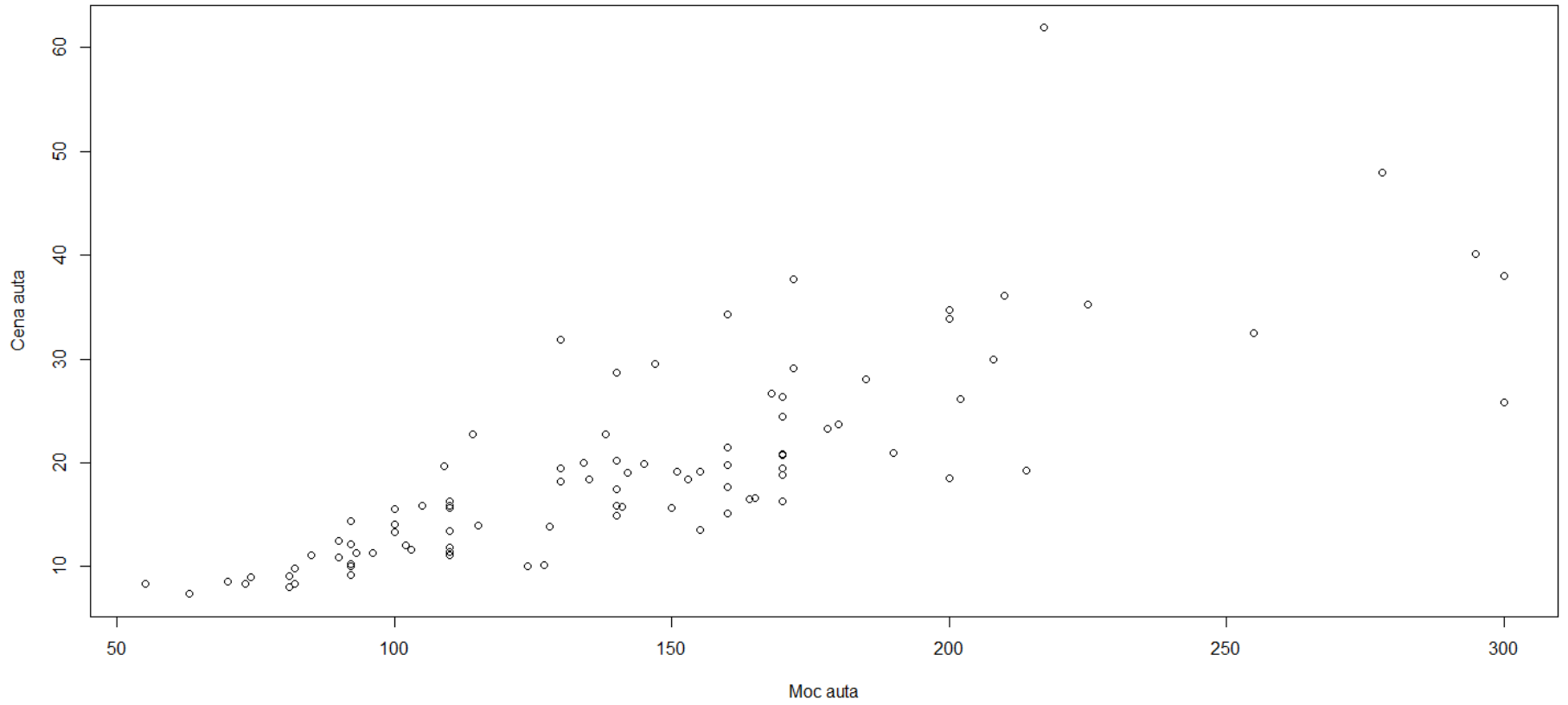
gdzie:

$$\begin{aligned} x_i' &= (x_{i,1}, x_{i,2}, ..., x_{i,p-1}) && = \text{zmienne objaśniające} \\ \beta' &= (\beta_0, \beta_1, ..., \beta_{p-1}) && = \text{współczynniki modelu regresji} \\ \epsilon_i & && = \text{błąd losowy} \end{aligned}$$

Regresja liniowa

- Reszty (***residuals***)
- Obserwacje odstające (***outliers***)
- Obserwacje istotne (***influence***)

Przykład w R: cena samochodu a jego moc



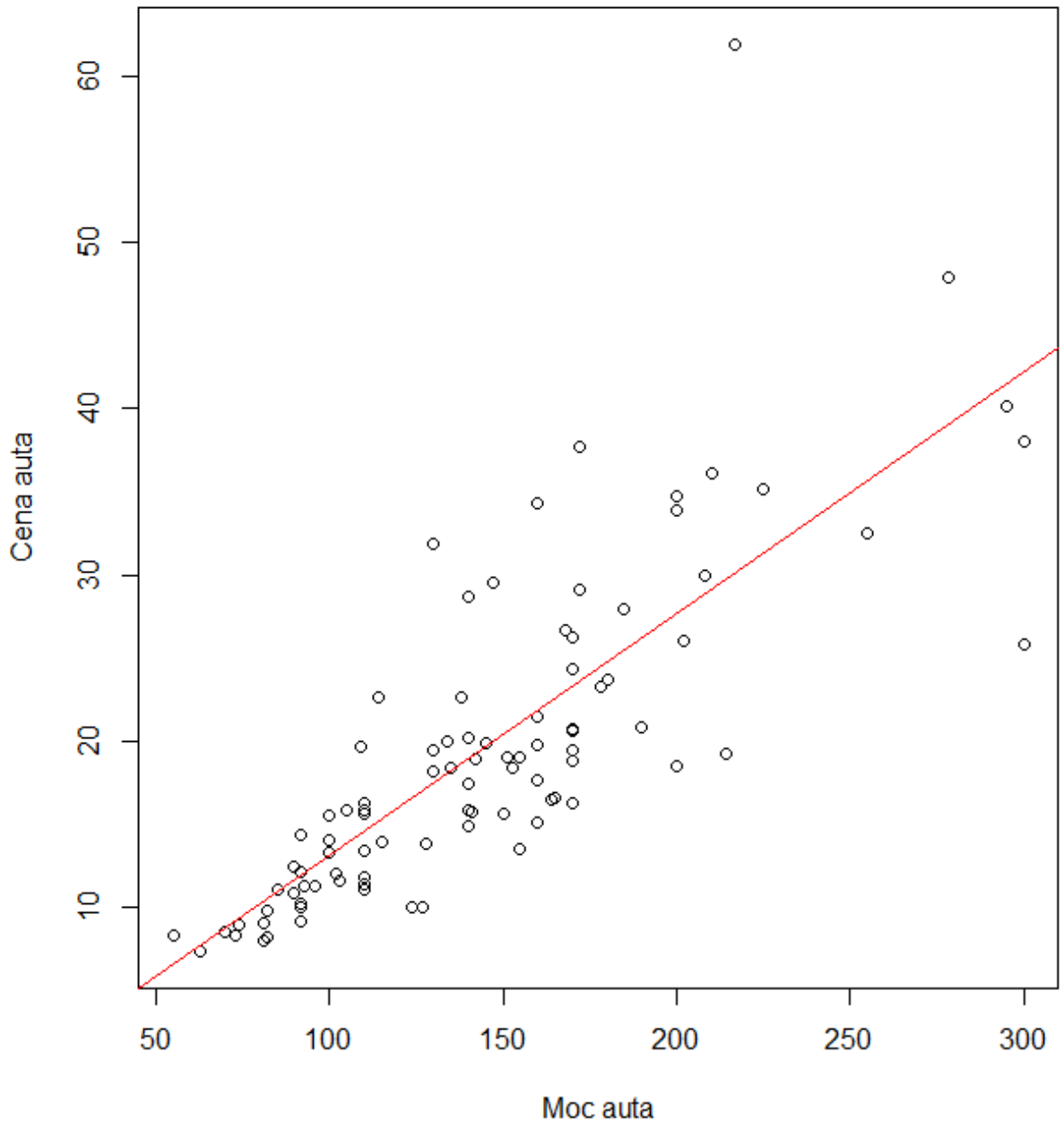
Przykład w R

```
Call:
lm(formula = Price ~ Horsepower, data = dane)

Residuals:
    Min      1Q  Median      3Q     Max
-16.413  -2.792  -0.821   1.803  31.753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3988     1.8200  -0.769    0.444
Horsepower    0.1454     0.0119  12.218   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.977 on 91 degrees of freedom
Multiple R-squared:  0.6213,     Adjusted R-squared:  0.6171
F-statistic: 149.3 on 1 and 91 DF,  p-value: < 2.2e-16
```

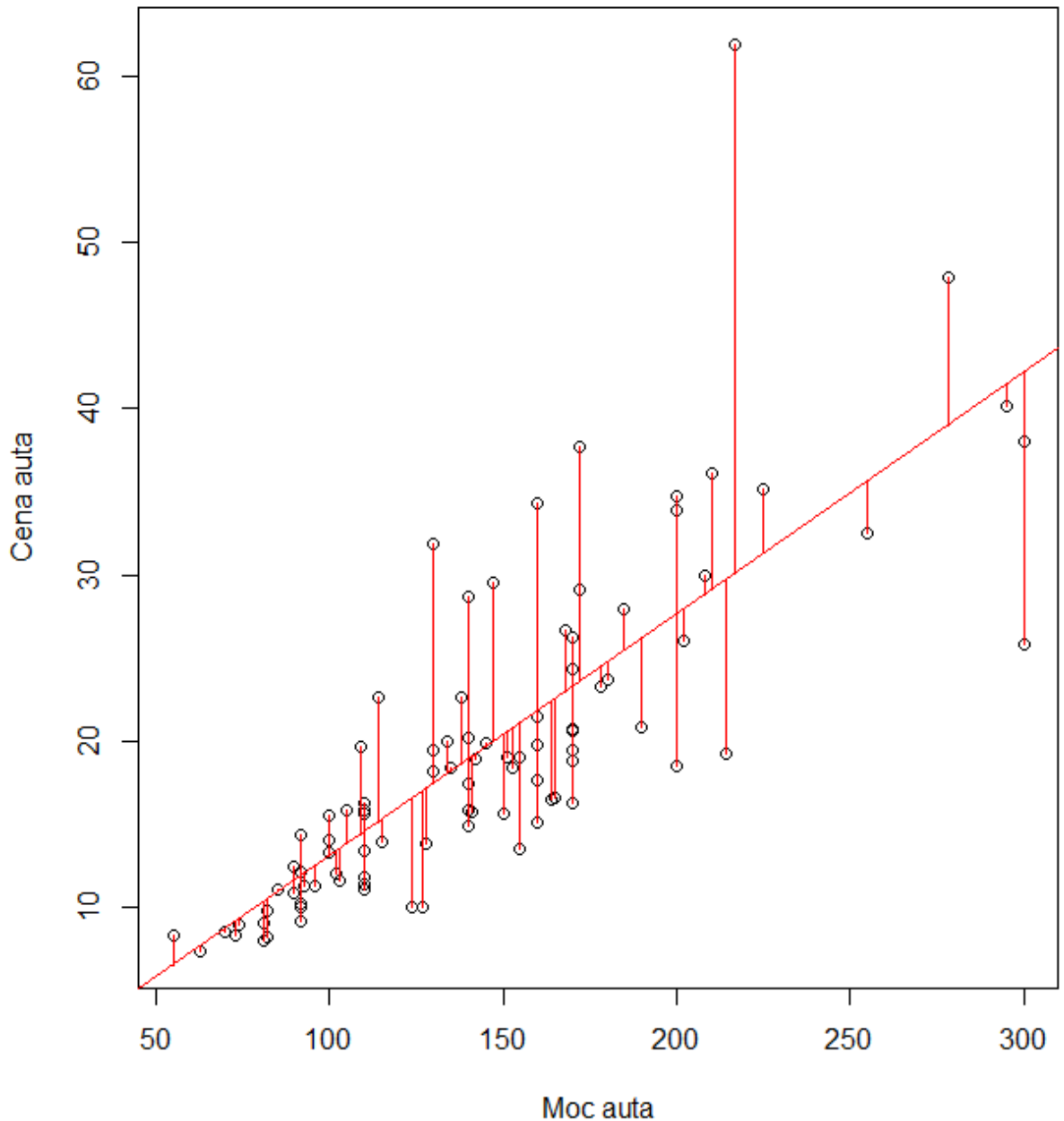
Przykład w R

```
Call:
lm(formula = Price ~ Horsepower, data = dane)

Residuals:
    Min      1Q  Median      3Q     Max
-16.413  -2.792  -0.821   1.803  31.753

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3988     1.8200  -0.769    0.444
Horsepower    0.1454     0.0119  12.218   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.977 on 91 degrees of freedom
Multiple R-squared:  0.6213,      Adjusted R-squared:  0.6171
F-statistic: 149.3 on 1 and 91 DF,  p-value: < 2.2e-16
```

Przykład w R

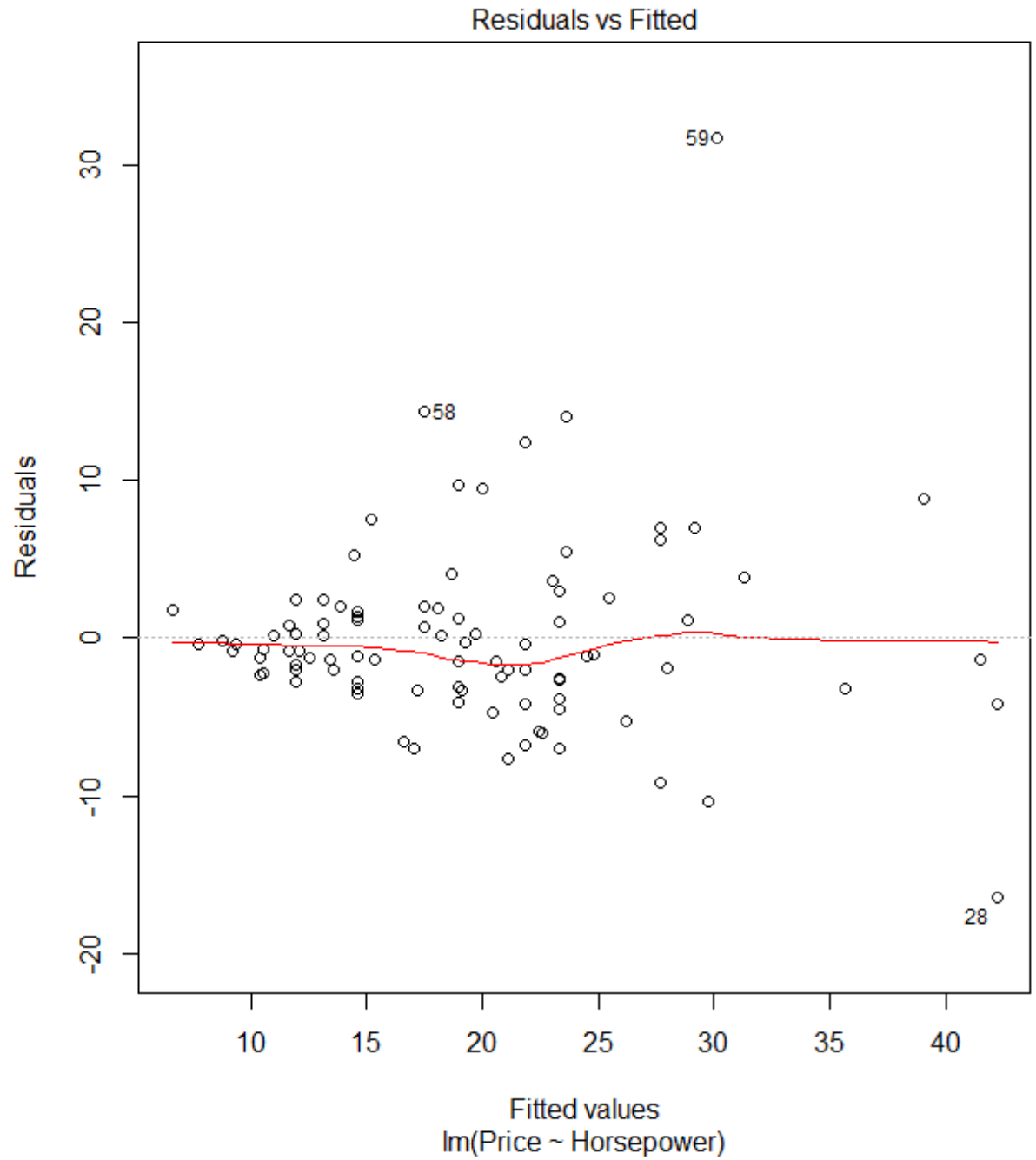


Przykład w R



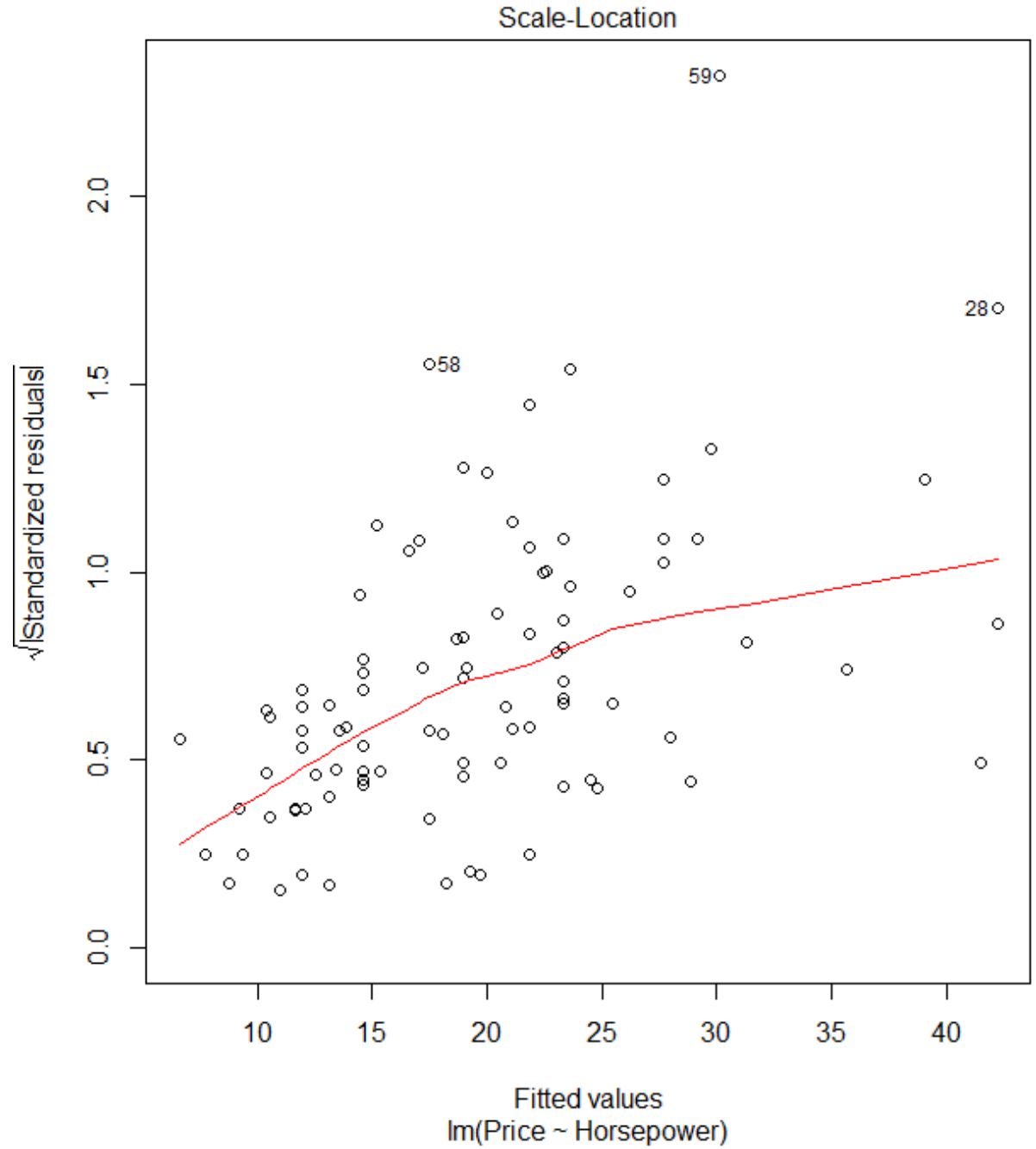
Diagnostyka modelu

Residuals vs fitted



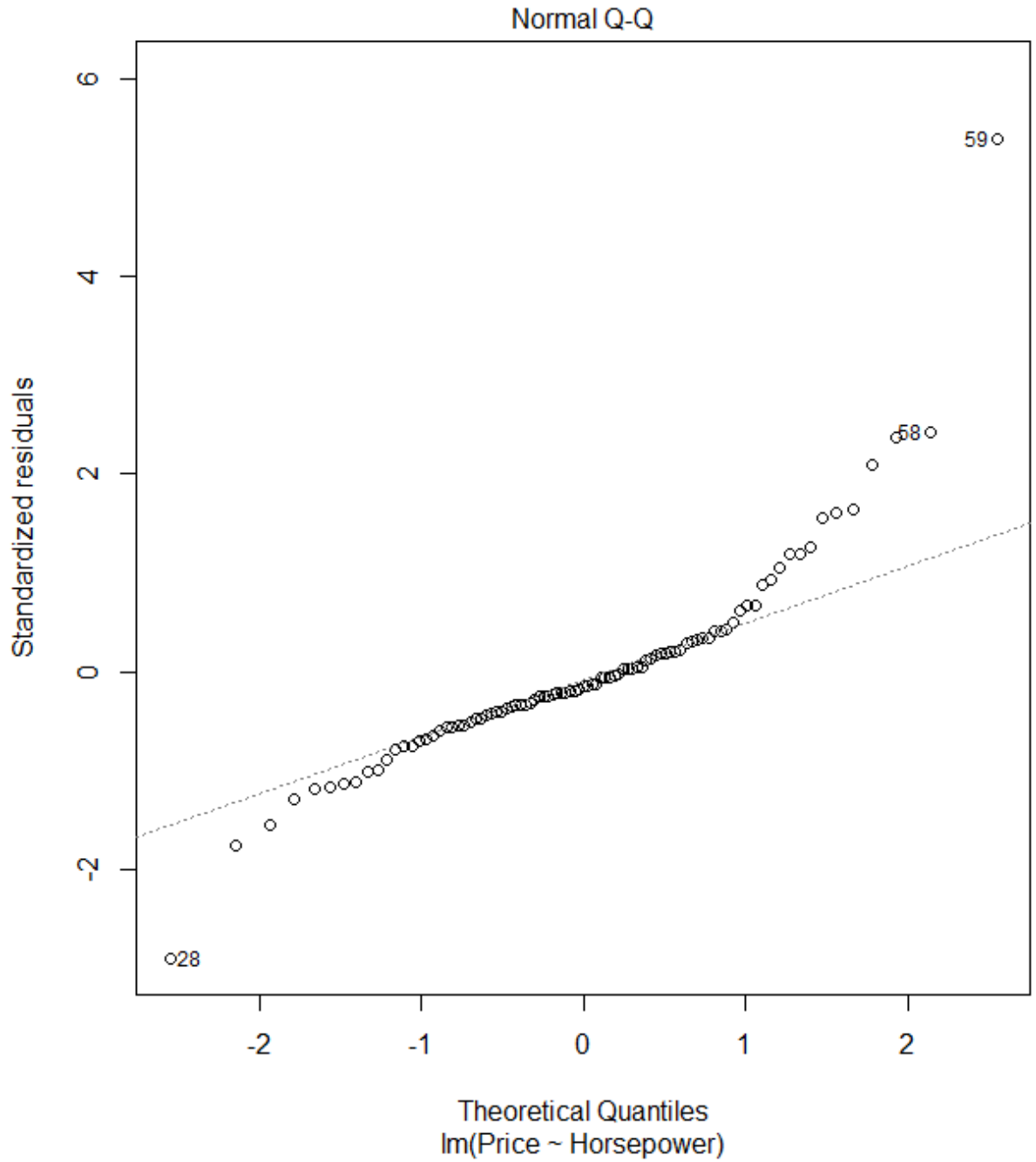
Diagnostyka modelu

Scale location



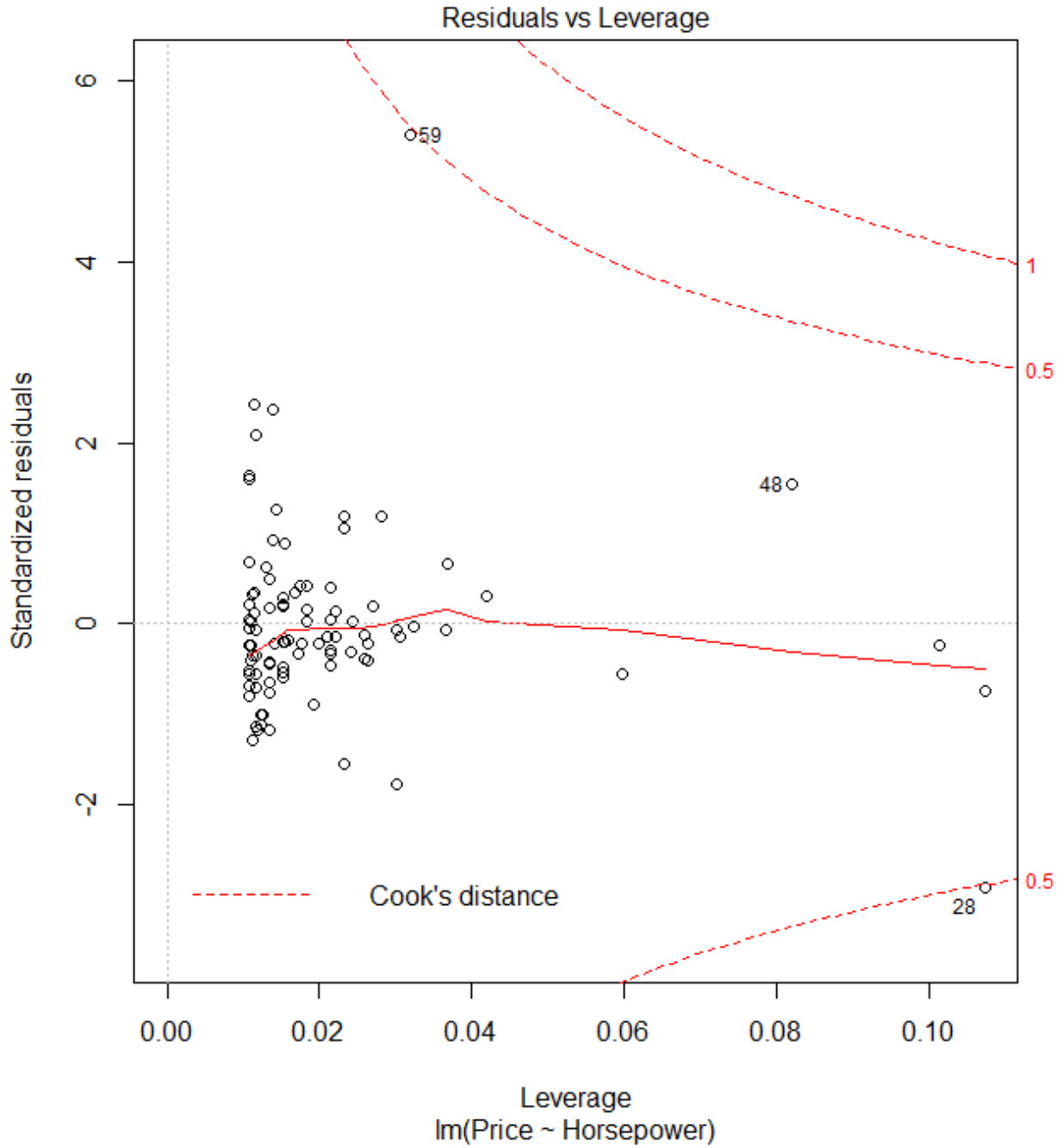
Diagnostyka modelu

Normal Q-Q



Diagnostyka modelu

Residuals vs Leverage



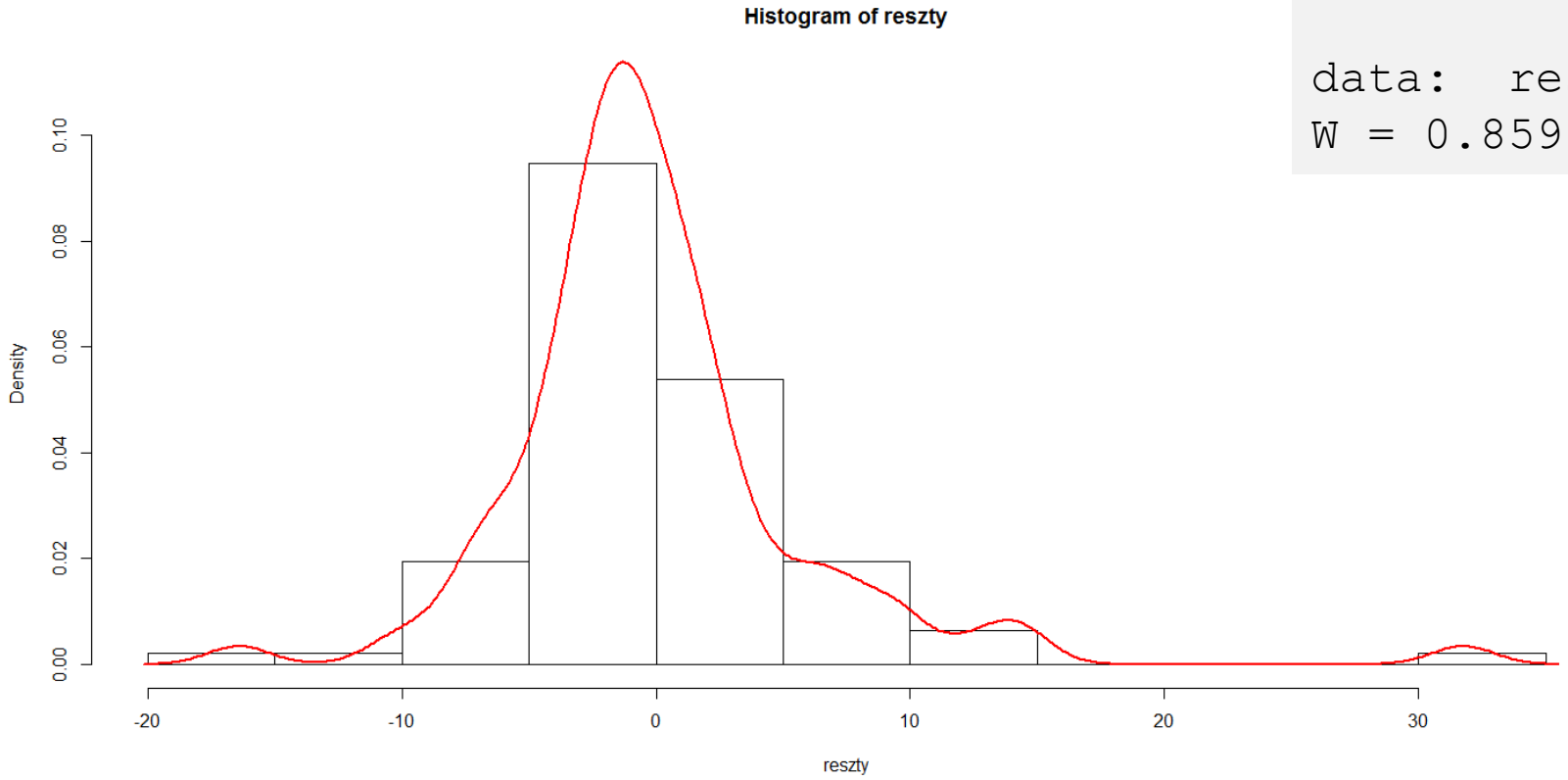
Przydatne funkcje

```
> coefficients(lm1)
(Intercept)  Horsepower
 -1.3987691   0.1453712
```

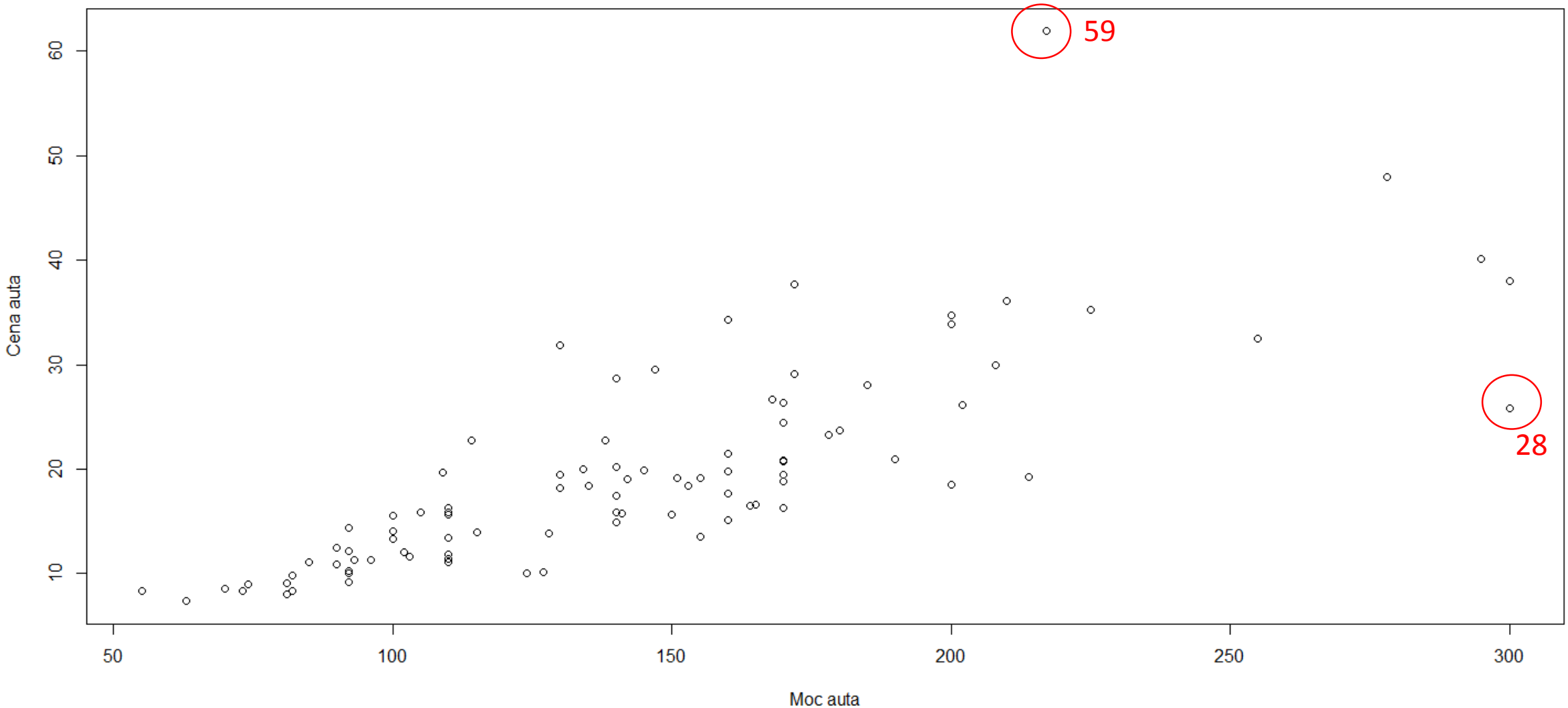
```
> residuals(lm1)
reszty = residuals(lm1)
hist(reszty, prob=TRUE, ylim=c(0,0.11))
lines(density(reszty), col="red", lwd=2)
```

```
> predict(lm1, dane)
```

Diagnostyka modelu

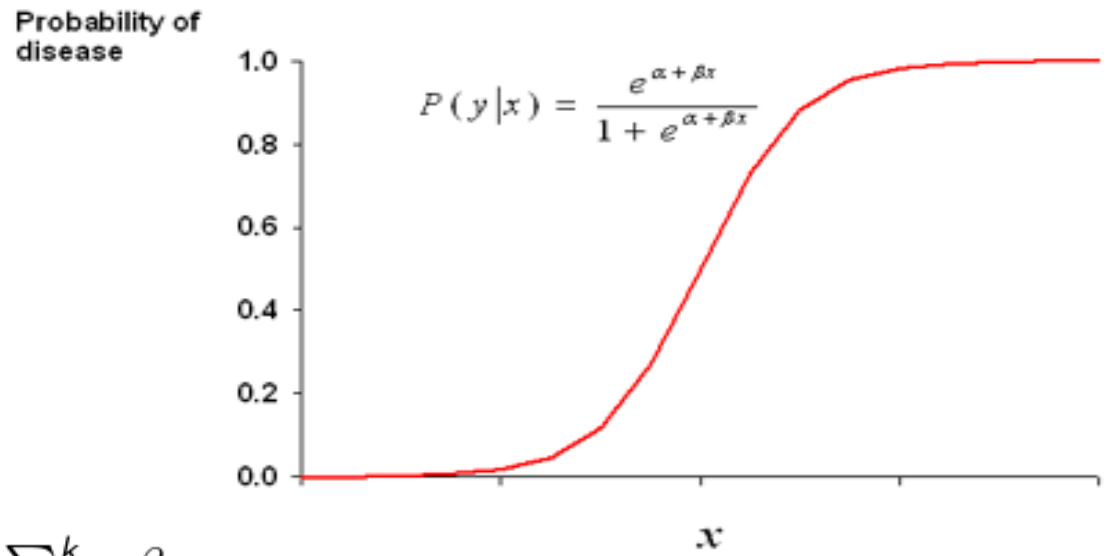



```
Shapiro-Wilk normality test

data:  reszty
W = 0.859, p-value = 6.113e-08
```


59
28
Cena auta
Moc auta
10
20
30
40
50
60
50
100
150
200
250
300

REGRESJA LOGISTYCZNA

Regresja logistyczna



$$P(Y = 1 \mid x_1, x_2, ..., x_k) = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}$$

gdzie:

$P(Y = 1 \mid x_1, x_2, ..., x_k)$ = prawdopodobieństwo warunkowe
$x_i' = (x_1, x_2, ..., x_k)$ = zmienne objaśniające
$\beta_i' = (\beta_0, \beta_1, ..., \beta_k)$ = współczynniki modelu regresji
e = stała Eulera, e $\approx$ 2.718

Regresja logistyczna: glm

```
fitlogit <- glm(formula = Churn ~ AccountLength + DayMins + IntlPlan + VMailPlan  +
IntlCalls + CustServCalls + EveCalls + NightCalls + VMailMessage, family = "binomial",
data = daneTrain)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.21781    0.89470   -5.83  5.5e-09 ***
AccountLength  0.00161    0.00259    0.62    0.534
DayMins        0.01329    0.00201    6.62  3.5e-11 ***
IntlPlan       1.70829    0.26012    6.57  5.1e-11 ***
VMailPlan     -2.46052    1.16908   -2.10    0.035 *
IntlCalls     -0.06824    0.04566   -1.49    0.135
CustServCalls  0.49943    0.07191    6.95  3.8e-12 ***
EveCalls       0.00726    0.00517    1.40    0.161
NightCalls    -0.00750    0.00523   -1.44    0.151
VMailMessage   0.04849    0.03610    1.34    0.179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regresja logistyczna: glm

```
fitlogit <- glm(formula = Churn ~ AccountLength + DayMins + IntlPlan + VMailPlan  +
IntlCalls + CustServCalls + EveCalls + NightCalls + VMailMessage, family = "binomial",
data = daneTrain)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.21781    0.89470   -5.83  5.5e-09 ***
AccountLength  0.00161    0.00259    0.62    0.534
DayMins        0.01329    0.00201    6.62  3.5e-11 ***
IntlPlan       1.70829    0.26012    6.57  5.1e-11 ***
VMailPlan     -2.46052    1.16908   -2.10    0.035 *
IntlCalls     -0.06824    0.04566   -1.49    0.135
CustServCalls  0.49943    0.07191    6.95  3.8e-12 ***
EveCalls       0.00726    0.00517    1.40    0.161
NightCalls    -0.00750    0.00523   -1.44    0.151
VMailMessage   0.04849    0.03610    1.34    0.179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dziękuję za uwagę