# Statistinės duomenų analizės praktinės užduotys

*2017*

10. **Tiesinė regresija**

(a) *Vieno kintamojo tiesinė regresija* (ČM II.5.2). Visuomeninis centras "Madam" kreipėsi į vyriausybę su prašymu visus valdininkus priversti lankyti jų centro organizuojamus džentelmeniškumo kursus. Kursų naudą pagrindė duomenimis, pateiktais lentelėje.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Treniruotės | 254 | 230 | 254 | 300 | 320 | 364 | 312 | 264 | 274 | 226 | 274 |
| Komplimentai | 124 | 108 | 85 | 152 | 140 | 198 | 182 | 125 | 130 | 95 | 171 |
| Treniruotės | 234 | 274 | 306 | 234 | 252 | 340 | 364 | 324 | 368 | 286 | 318 |
| Komplimentai | 102 | 115 | 109 | 115 | 134 | 213 | 155 | 188 | 204 | 85 | 148 |
| Treniruotės | 216 | 350 | 216 | 358 | 222 | 374 | 222 | 230 | 360 | 336 | |
| Komplimentai | 106 | 155 | 73 | 179 | 118 | 159 | 79 | 74 | 180 | 126 | |

```
dat.a <- matrix(c(254, 124,  230, 108,  254,  85,  300, 152,  320, 140,
                  364, 198,  312, 182,  264, 125,  274, 130,  226,  95,
                  274, 171,  234, 102,  274, 115,  306, 109,  234, 115,
                  252, 134,  340, 213,  364, 155,  324, 188,  368, 204,
                  286,  85,  318, 148,  216, 106,  350, 155,  216,  73,
                  358, 179,  222, 118,  374, 159,  222,  79,  230,  74,
                  360, 180,  336, 126),
                byrow = TRUE,
                nrow = 32, ncol = 2)

dat.a<-as.data.frame(dat.a)
colnames(dat.a) <- c("Treniruotes", "Komplimentai")
dat.a
```

```
##    Treniruotes Komplimentai
## 1          254          124
## 2          230          108
## 3          254           85
## 4          300          152
## 5          320          140
## 6          364          198
## 7          312          182
## 8          264          125
## 9          274          130
## 10         226           95
## 11         274          171
## 12         234          102
## 13         274          115
## 14         306          109
## 15         234          115
## 16         252          134
## 17         340          213
## 18         364          155
## 19         324          188
## 20         368          204
## 21         286           85
```

```
## 22         318          148
## 23         216          106
## 24         350          155
## 25         216           73
## 26         358          179
## 27         222          118
## 28         374          159
## 29         222           79
## 30         230           74
## 31         360          180
## 32         336          126
```

Joje užfiksuota, kiek kartų kiekvienas vyras kursuose treniravosi buti džentelmenu ($x$) ir kiek po to per menesį savo žmonai viešai pasakė komplimentų (skaičiavo uošvė). Ištirkite, ar tinka tiesinės regresijos modelis, ir padarykite prognozę, kiek komplimentų pasakys vyras, treniravęsis 267 kartus. Sudarykite 95% prognozės ir vidurkio pasikliovimo intervalus.

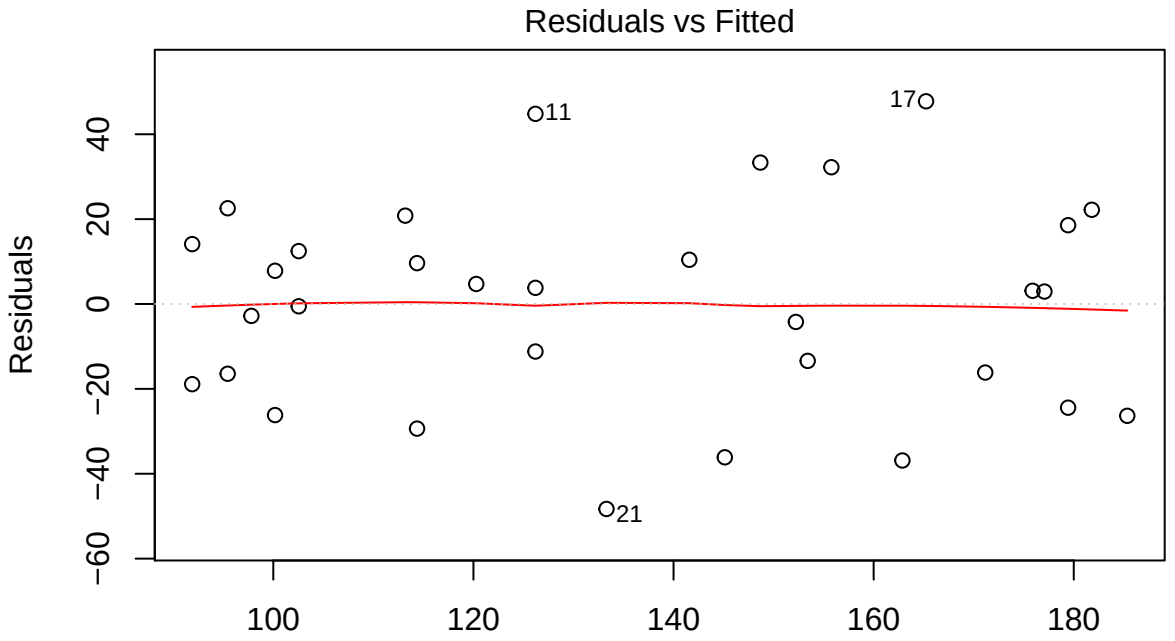
```
modelis.a <- lm(Komplimentai~Treniruotes, data = dat.a)
summary(modelis.a)
```

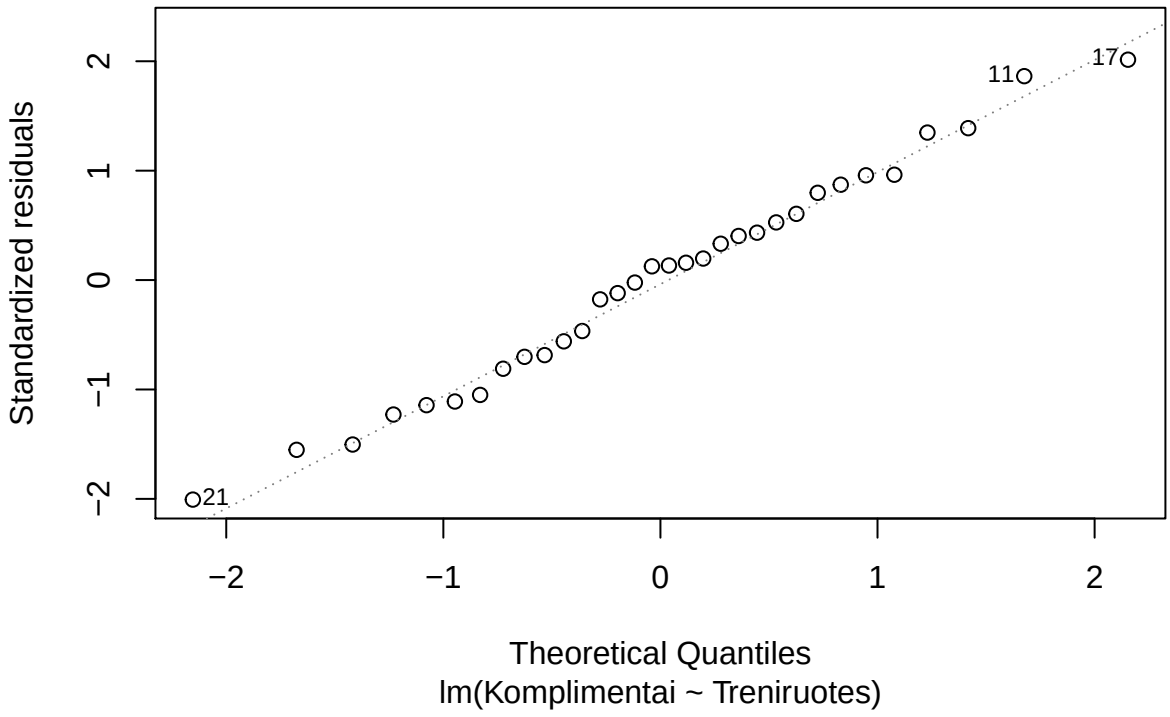
```
##
## Call:
## lm(formula = Komplimentai ~ Treniruotes, data = dat.a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.296 -17.052   3.023  15.223  47.762
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.87554   24.13549  -1.486    0.148
## Treniruotes   0.59151    0.08209   7.205  5.1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.46 on 30 degrees of freedom
## Multiple R-squared:  0.6338, Adjusted R-squared:  0.6216
## F-statistic: 51.92 on 1 and 30 DF,  p-value: 5.097e-08
```

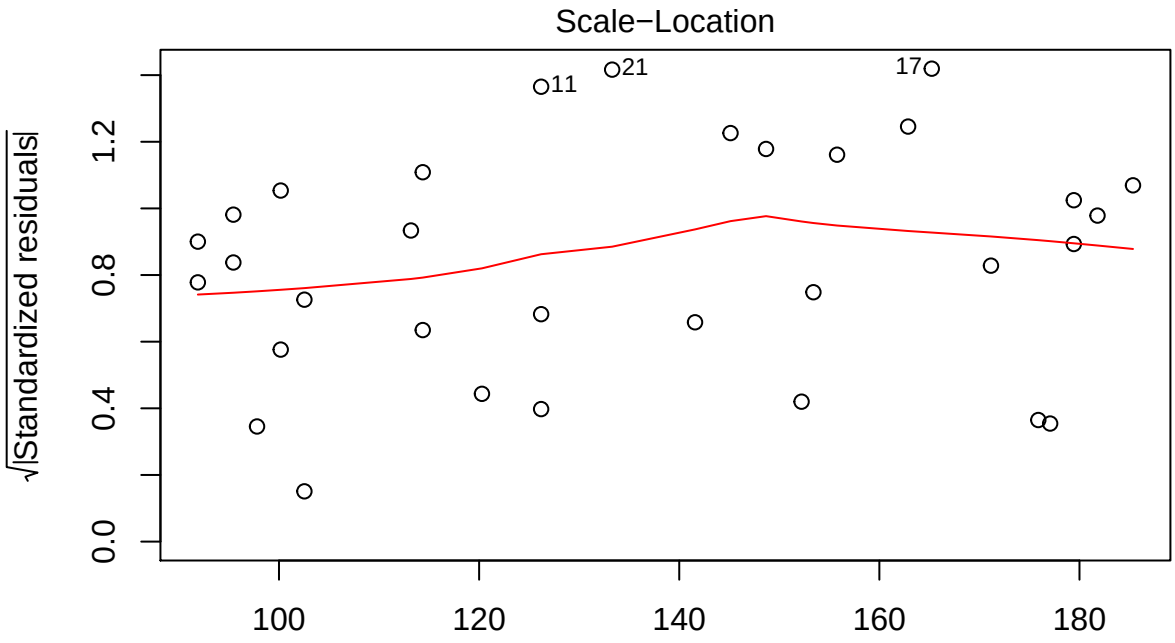
```
ats <- data.frame(267)
colnames(ats) <- c("Treniruotes")
predict(modelis.a, ats, interval="confidence")
```

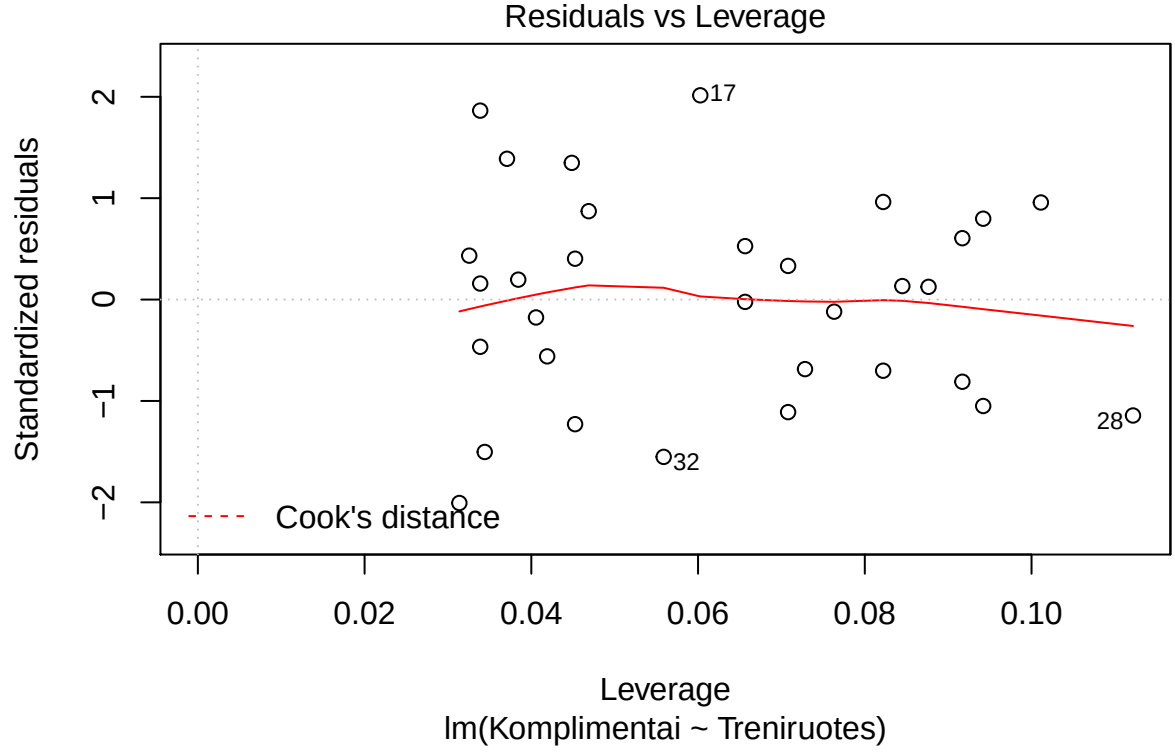
```
##        fit      lwr      upr
## 1 122.0577 112.4724 131.6429
```

```
plot(modelis.a)
```

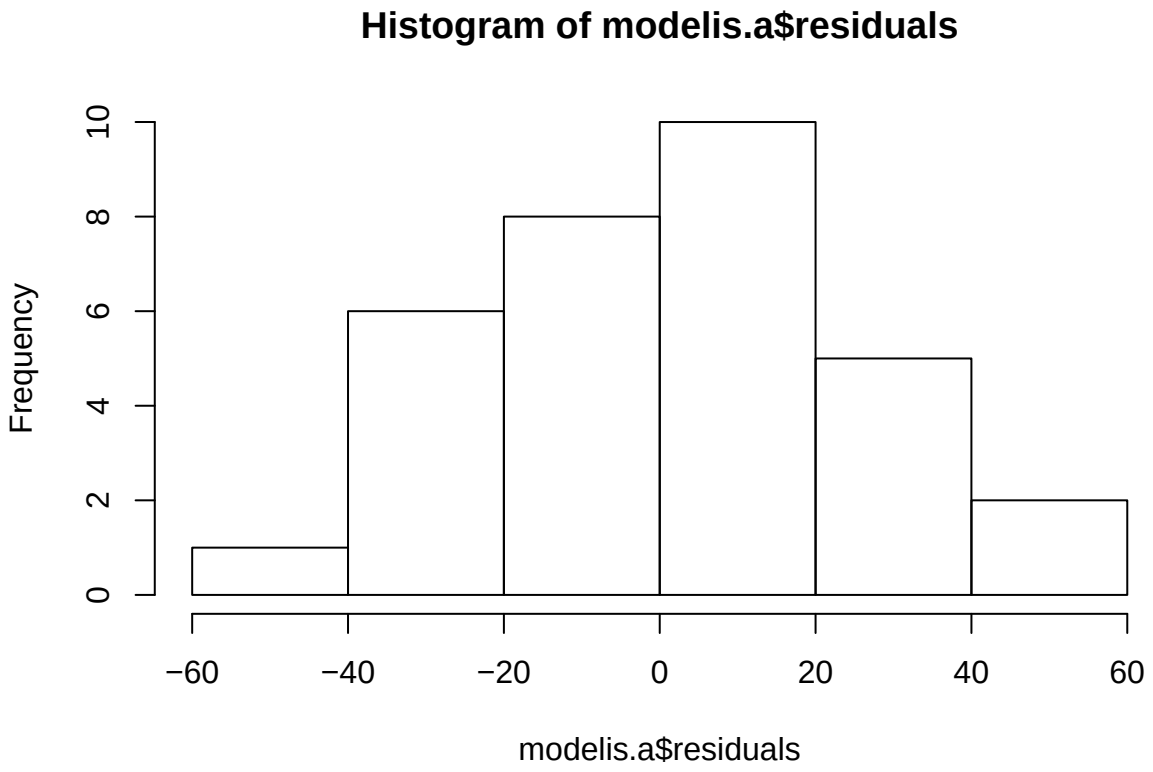

Residuals vs Fitted
Residuals
Fitted values
lm(Komplimentai ~ Treniruotes)
11
17
21
−60
−40
−20
0
20
40
100
120
140
160
180


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Komplimentai ~ Treniruotes)
11
17
21
−2
−1
0
1
2




```
hist(modelis.a$residuals)
```




```
confint(modelis.a, 'Treniruotes', level = 0.95)
```

```
##                 2.5 %    97.5 %
## Treniruotes 0.4238559 0.7591642
```

(b) *Vieno kintamojo tiesinė regresija* (ČM II.5.3). Paplūdimio gelbėtojų tarnyba visą mėnesį fiksavo vandens temperatūrą ir maksimalų besimaudančiųjų skaičių. Duomenys pateikti lentelėje.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vandens temperatūra | 17 | 18 | 16 | 18 | 16 | 18 | 14 | 15 | 19 | 12 | 12 | 14 |
| Mauduolių skaičius | 79 | 83 | 78 | 82 | 78 | 81 | 74 | 76 | 84 | 71 | 72 | 73 |
| Vandens temperatūra | 20 | 14 | 15 | 15 | 20 | 16 | 18 | 21 | 17 | 17 | 16 | 12 |
| Mauduolių skaičius | 85 | 76 | 76 | 77 | 86 | 79 | 82 | 88 | 81 | 82 | 79 | 120 |

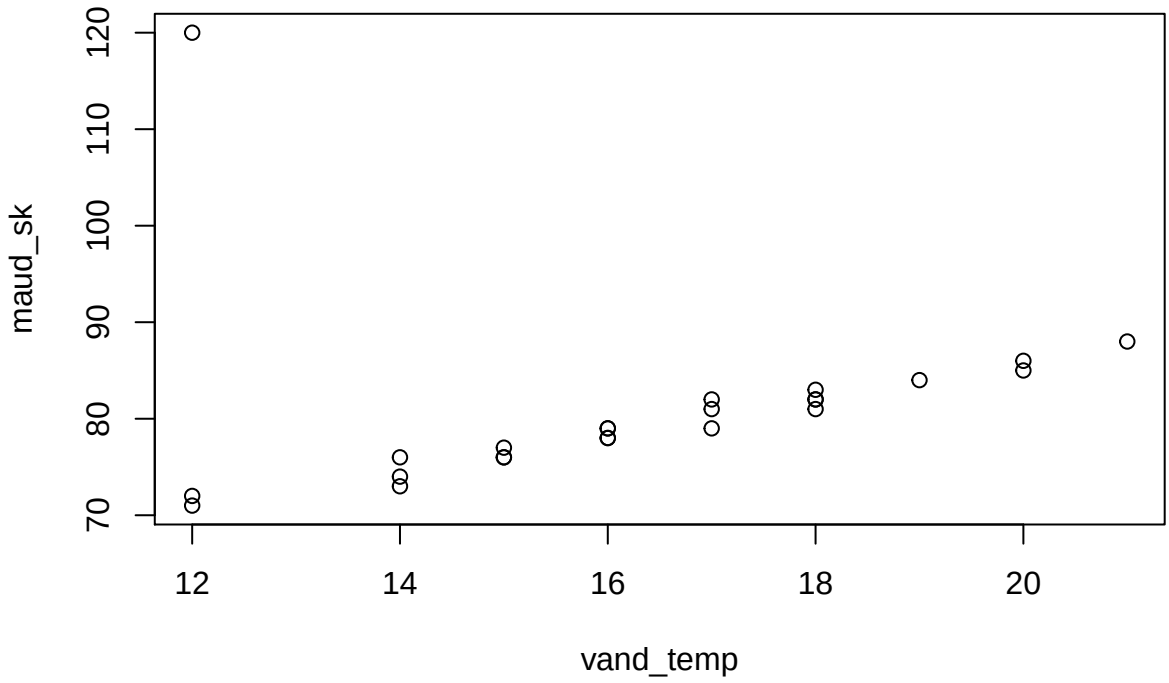
```
dat.b <- matrix(c(17, 79,  18, 83,  16, 78,  18, 82,  16, 78,
                  18, 81,  14, 74,  15, 76,  19, 84,  12, 71,
                  12, 72,  14, 73,  20, 85,  14, 76,  15, 76,
                  15, 77,  20, 86,  16, 79,  18, 82,  21, 88,
                  17, 81,  17, 82,  16, 79,  12, 120),
                byrow = TRUE,
                nrow = 24, ncol = 2)

dat.b<-as.data.frame(dat.b)
colnames(dat.b) <- c("vand_temp", "maud_sk")
dat.b
```

```
##    vand_temp maud_sk
## 1         17      79
## 2         18      83
## 3         16      78
## 4         18      82
```

```
## 5          16       78
## 6          18       81
## 7          14       74
## 8          15       76
## 9          19       84
## 10         12       71
## 11         12       72
## 12         14       73
## 13         20       85
## 14         14       76
## 15         15       76
## 16         15       77
## 17         20       86
## 18         16       79
## 19         18       82
## 20         21       88
## 21         17       81
## 22         17       82
## 23         16       79
## 24         12      120
```

Ištirkite, ar vandens temperatura įtakoja mauduolių skaičių. Reikšmingumo lygmuo $\alpha = 0.05$. Ką galima pasakyti apie regresijos modelį, pašalinus išskirtis?

```
plot(dat.b)
```

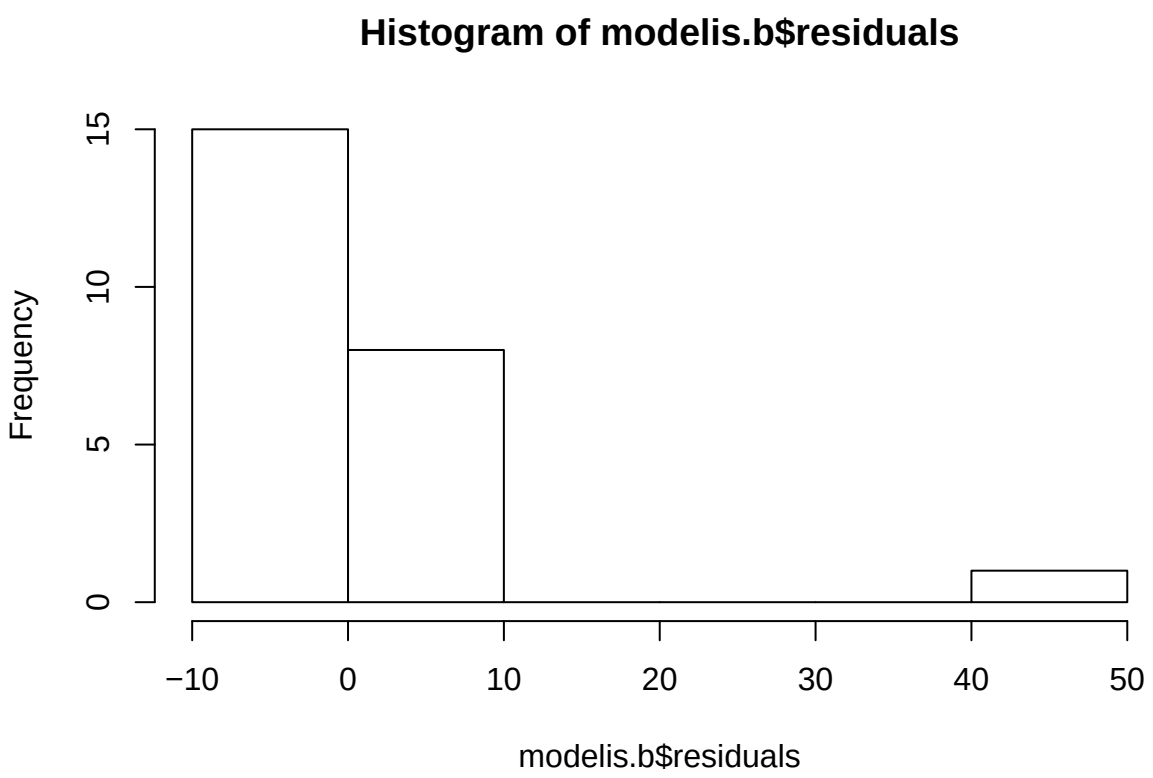



```
modelis.b <- lm(maud_sk~vand_temp, data = dat.b)
summary(modelis.b)
```

```
##
## Call:
## lm(formula = maud_sk ~ vand_temp, data = dat.b)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -8.133 -4.077 -1.812  0.914 40.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  74.0950    12.9997   5.700 9.86e-06 ***
## vand_temp     0.4198     0.7909   0.531    0.601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.573 on 22 degrees of freedom
## Multiple R-squared:  0.01264,    Adjusted R-squared:  -0.03224
## F-statistic: 0.2817 on 1 and 22 DF,  p-value: 0.6009
```

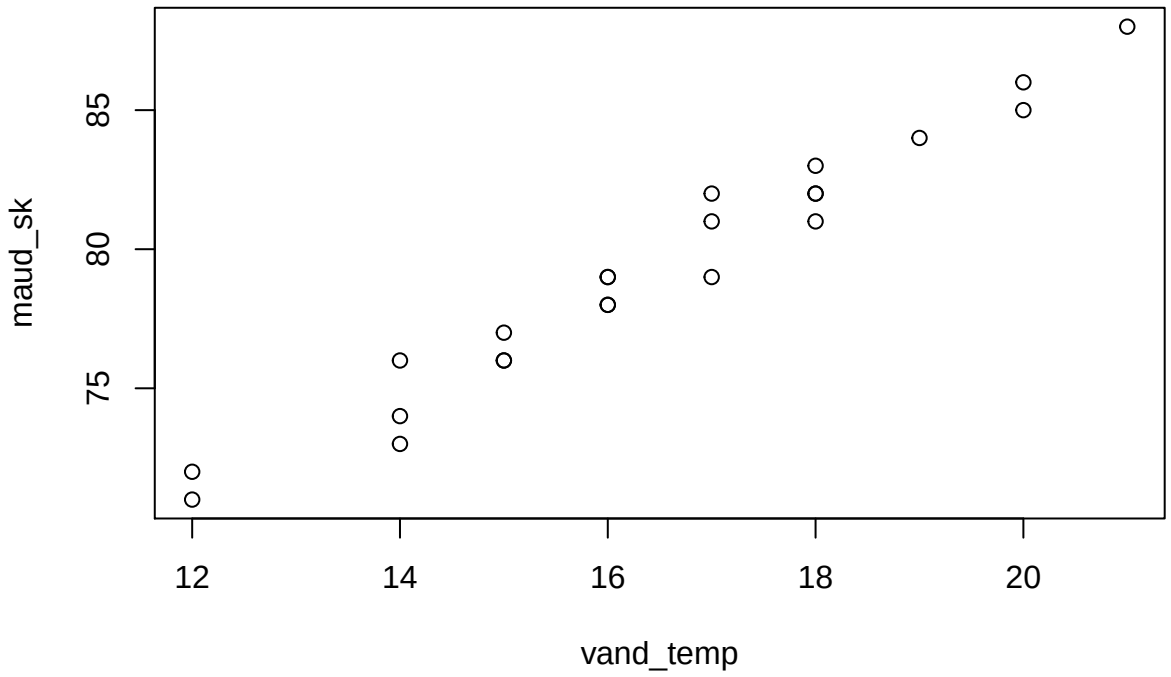
```
hist(modelis.b$residuals)
```




```
confint(modelis.b, 'vand_temp', level = 0.95)
```

```
##                2.5 %   97.5 %
## vand_temp -1.220413 2.060003
```

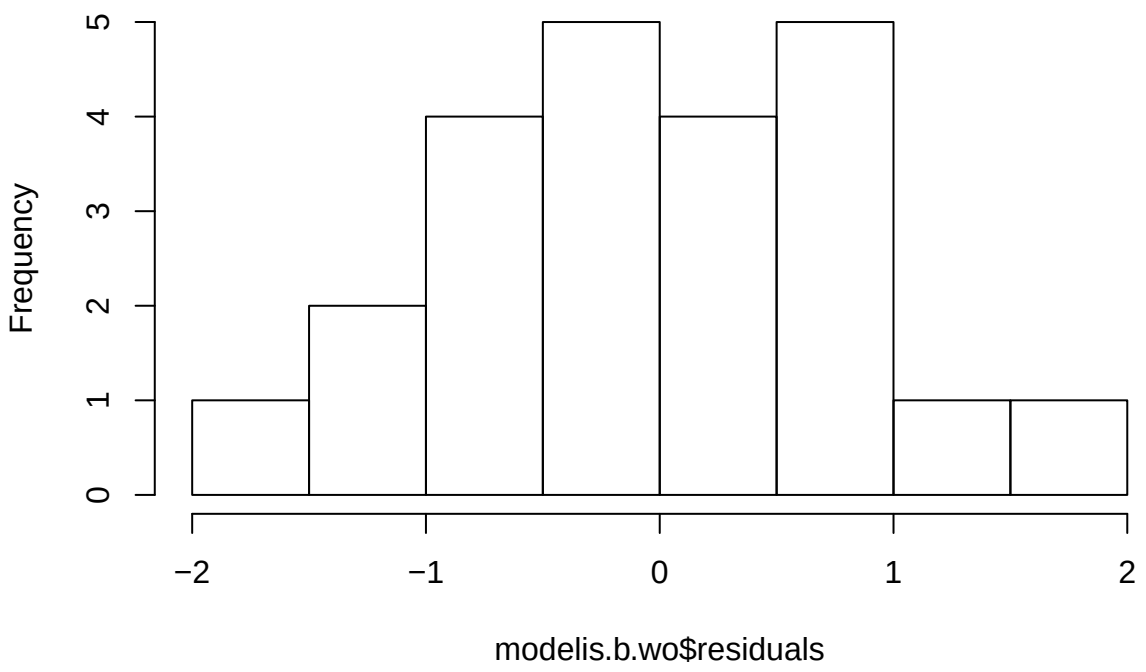
```
# without outliers
dat.b.wo <- dat.b[which(dat.b$maud_sk < 120),]
plot(dat.b.wo)
```




```
modelis.b.wo <- lm(maud_sk~vand_temp, data = dat.b.wo)
summary(modelis.b.wo)
```

```
##
## Call:
## lm(formula = maud_sk ~ vand_temp, data = dat.b.wo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73842 -0.57800 -0.05926  0.58243  1.74285
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.98433    1.27996   38.27   <2e-16 ***
## vand_temp    1.83958    0.07709   23.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.871 on 21 degrees of freedom
## Multiple R-squared:  0.9644, Adjusted R-squared:  0.9627
## F-statistic: 569.4 on 1 and 21 DF,  p-value: < 2.2e-16
```

```
hist(modelis.b.wo$residuals)
```




```
confint(modelis.b.wo, 'vand_temp', level = 0.95)
```

```
##              2.5 %   97.5 %
## vand_temp 1.679254 1.999901
```

(c) *Kelių kintamųjų tiesinė regresija.* (ČM II.6.1). Policijos komisaras pastebėjo, kad po kiekvienos stambesnės privatizacijos kai kurie valdininkai gauna iš *nežinomos tetos* dovanų (palikimus). Duomenys apie dovanų vertę (tūkst. dolerių), privatizuotų objektų kainą (mln.eurų) ir konkurse dalyvavusių firmų skaičių pateikti lentelėje.

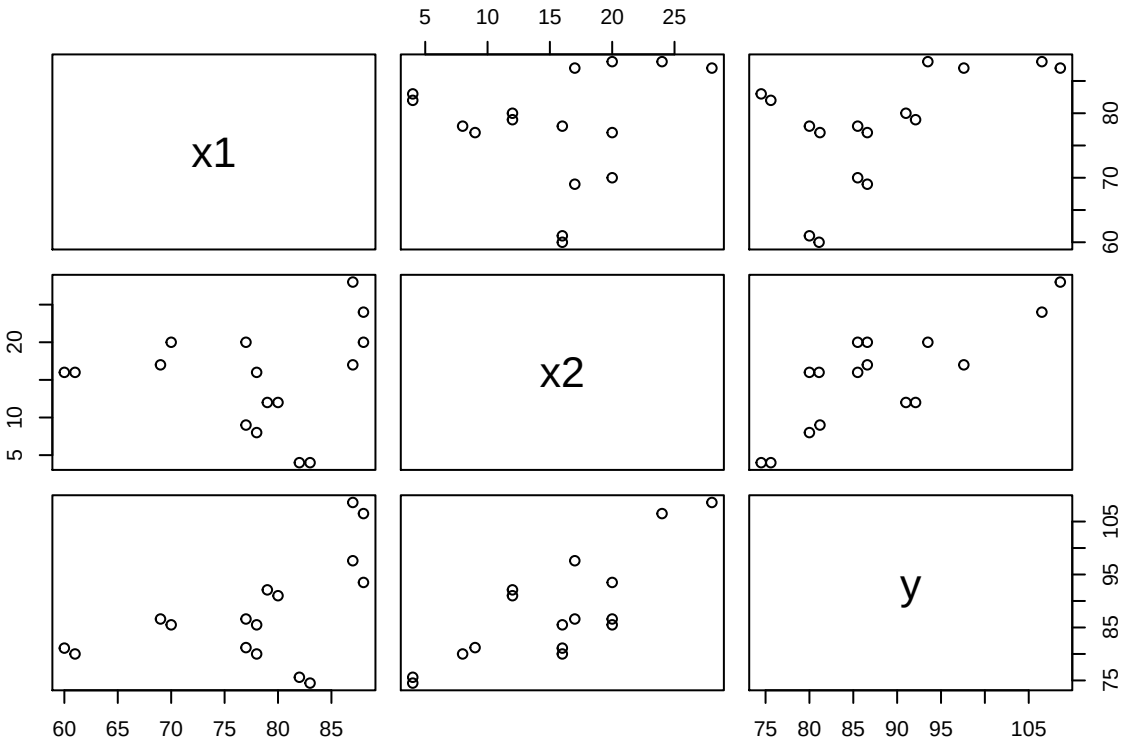
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Vertė ($x_1$) | 88 | 83 | 88 | 78 | 70 | 80 | 61 | 78 |
| Firmų skaičius ($x_2$) | 24 | 4 | 20 | 8 | 20 | 12 | 16 | 16 |
| Dovana ($Y$) | 106,5 | 74,5 | 93,5 | 80 | 85,5 | 91 | 80 | 85,5 |
| Vertė ($x_1$) | 87 | 82 | 87 | 77 | 69 | 79 | 60 | 77 |
| Firmų skaičius ($x_2$) | 28 | 4 | 17 | 9 | 17 | 12 | 16 | 20 |
| Dovana ($Y$) | 108,6 | 75,6 | 97,6 | 81,2 | 86,6 | 92,1 | 81,1 | 86,6 |

```
dat.c <- matrix(c(88, 24, 106.5,  83,  4, 74.5,  88, 20,  93.5,
                  78,  8,  80.0,  70, 20, 85.5,  80, 12,  91.0,
                  61, 16,  80.0,  78, 16, 85.5,  87, 28, 108.6,
                  82,  4,  75.6,  87, 17, 97.6,  77,  9,  81.2,
                  69, 17,  86.6,  79, 12, 92.1,  60, 16,  81.1,
                  77, 20,  86.6),
                byrow = TRUE,
                nrow = 16, ncol = 3)

dat.c<-as.data.frame(dat.c)
colnames(dat.c) <- c("x1", "x2", "y")
dat.c
```

```
##    x1 x2     y
## 1  88 24 106.5
## 2  83  4  74.5
## 3  88 20  93.5
## 4  78  8  80.0
## 5  70 20  85.5
## 6  80 12  91.0
## 7  61 16  80.0
## 8  78 16  85.5
## 9  87 28 108.6
## 10 82  4  75.6
## 11 87 17  97.6
## 12 77  9  81.2
## 13 69 17  86.6
## 14 79 12  92.1
## 15 60 16  81.1
## 16 77 20  86.6
```

Kokią tetos dovaną prognozuotumete valdininkui, jei objektas privatizuotas už 90 mln. eurų, o konkurse dalyvavo 10 firmų? Ar šiems duomenims tinka tiesinės regresijos modelis?

```
plot(dat.c)
```

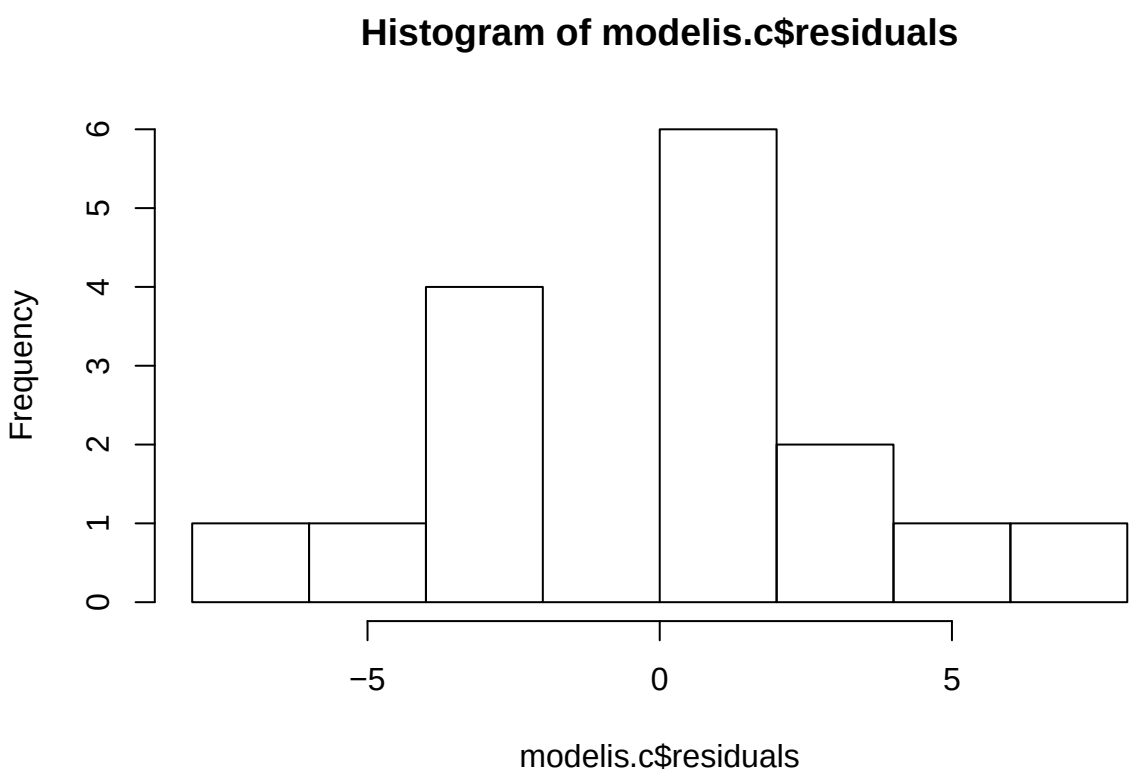



```
modelis.c <- lm(y~x1+x2, data = dat.c)
summary(modelis.c)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = dat.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.2581 -3.4671  0.3923  2.0151  7.1354
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.5270     9.4639   3.120 0.008128 **
## x1            0.5317     0.1197   4.441 0.000665 ***
## x2            1.1196     0.1568   7.142 7.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.075 on 13 degrees of freedom
## Multiple R-squared:  0.855,  Adjusted R-squared:  0.8327
## F-statistic: 38.34 on 2 and 13 DF,  p-value: 3.534e-06
```

```
hist(modelis.c$residuals)
```




```
ats <- data.frame(90, 10)
colnames(ats)<-c("x1", "x2")
predict(modelis.c, ats)
```

```
##        1
## 88.57384
```

(d) *Kelių kintamųjų tiesinė regresija* (ČM II.6.2). Sporto apžvalgininkas tiria, kaip krepšinio komandos laimetų rungtynių procentas $Y$ priklauso nuo komandos biudžeto $x_1$ (mln.litų), vidutinio per rungtynes pelnyto taškų skaičiaus $x_2$ ir tritaškių pataikymo procento $x_3$. Duomenys apie 16 krepšinio komandų pateikti lentelėje.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Pergalės ($Y$) | 91 | 33 | 73 | 43 | 53 | 63 | 43 | 53 |
| Biudžetas ($x_1$) | 6 | 7 | 6,1 | 8 | 9,6 | 7,6 | 11,4 | 8 |
| Taškai ($x_2$) | 85 | 74 | 83 | 77 | 78 | 81 | 80 | 79 |

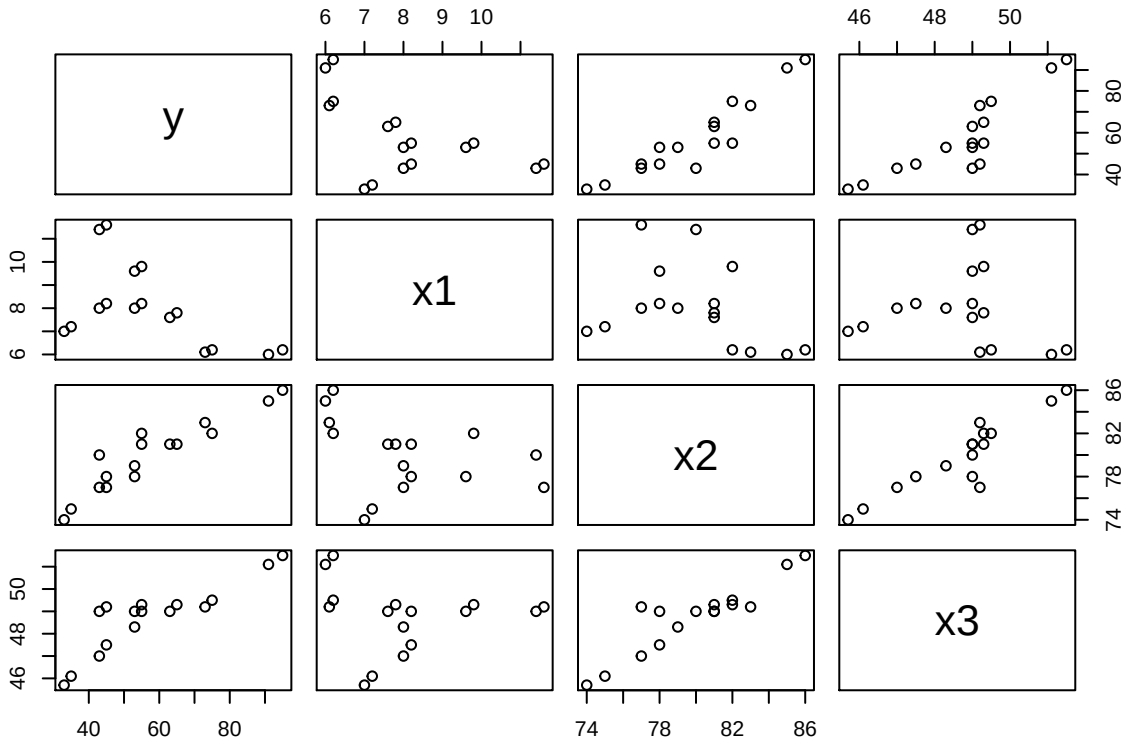
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tritaškiai ($x_3$) | 51,1 | 45,7 | 49,2 | 47 | 49 | 49 | 49 | 48,3 |
| Pergalės ($Y$) | 95 | 35 | 75 | 45 | 55 | 65 | 45 | 55 |
| Biudžetas ($x_1$) | 6,2 | 7,2 | 6,2 | 8,2 | 9,8 | 7,8 | 11,6 | 8,2 |
| Taškai ($x_2$) | 86 | 75 | 82 | 78 | 82 | 81 | 77 | 81 |
| Tritaškiai ($x_3$) | 51,5 | 46,1 | 49,5 | 47,5 | 49,3 | 49,3 | 49,2 | 49 |

```
dat.d <- matrix(c(91,  6.0, 85, 51.1,  33, 7.0, 74, 45.7,  73,  6.1, 83, 49.2,
                  43,  8.0, 77, 47.0,  53, 9.6, 78, 49.0,  63,  7.6, 81, 49.0,
                  43, 11.4, 80, 49.0,  53, 8.0, 79, 48.3,  95,  6.2, 86, 51.5,
                  35,  7.2, 75, 46.1,  75, 6.2, 82, 49.5,  45,  8.2, 78, 47.5,
                  55,  9.8, 82, 49.3,  65, 7.8, 81, 49.3,  45, 11.6, 77, 49.2,
                  55,  8.2, 81, 49.0),
                byrow = TRUE,
                nrow = 16, ncol = 4)

dat.d<-as.data.frame(dat.d)
colnames(dat.d) <- c("y", "x1", "x2", "x3")
dat.d
```

```
##     y   x1 x2   x3
## 1  91  6.0 85 51.1
## 2  33  7.0 74 45.7
## 3  73  6.1 83 49.2
## 4  43  8.0 77 47.0
## 5  53  9.6 78 49.0
## 6  63  7.6 81 49.0
## 7  43 11.4 80 49.0
## 8  53  8.0 79 48.3
## 9  95  6.2 86 51.5
## 10 35  7.2 75 46.1
## 11 75  6.2 82 49.5
## 12 45  8.2 78 47.5
## 13 55  9.8 82 49.3
## 14 65  7.8 81 49.3
## 15 45 11.6 77 49.2
## 16 55  8.2 81 49.0
```

Ar tiesinės regresijos modelis tinka? Ar visi kintamieji jame reikalingi?

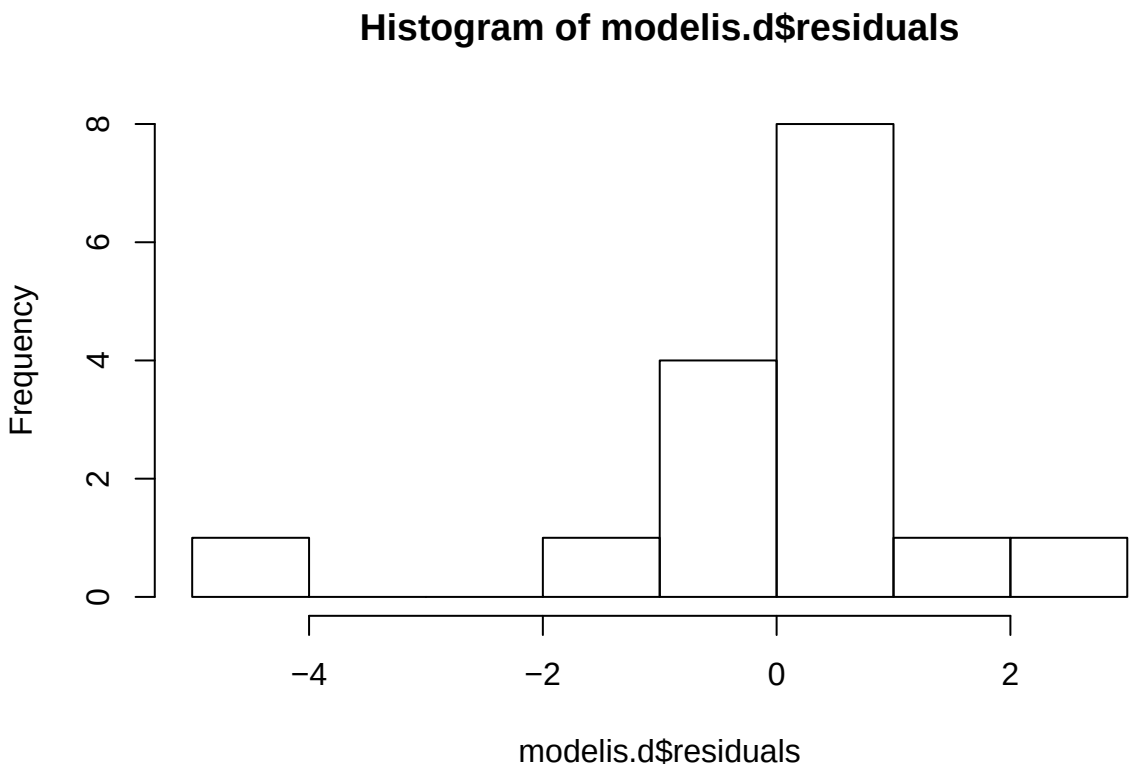
```
plot(dat.d)
```




```
modelis.d <- lm(y~., data = dat.d)
summary(modelis.d)
```

```
##
## Call:
## lm(formula = y ~ ., data = dat.d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4667 -0.2674  0.3371  0.4904  2.2461
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -384.4342    13.6442 -28.176 2.48e-12 ***
## x1            -5.1310     0.3190 -16.087 1.74e-09 ***
## x2            -0.1393     0.4045  -0.344    0.736
## x3            10.1482     0.8216  12.352 3.50e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.573 on 12 degrees of freedom
## Multiple R-squared:  0.9941, Adjusted R-squared:  0.9927
## F-statistic: 676.7 on 3 and 12 DF,  p-value: 1.204e-13
```

```
hist(modelis.d$residuals)
```




Kintamasis $x_2$ nereikalingas.

---

Padaryta su R version 3.4.2 (2017-09-28), x86_64-pc-linux-gnu.