NEX: Homework Assignment 01

Belov, Neoral, Sahan, Shulga

October 30, 2018

Data

In this assigment we use a data set "experiment_data". The data contain the total number of dots (number of hits) inside circles with different diameters which different testees (operators) were able to put in during the time interval of 10 seconds. The data were collected from the results of four students. The dataset consists of 36 observations of 4 variables.

- BLOCK - testees (operators);
- HITS_SUM - total number of hits;
- DIAMETER - the diameter of the circle in [cm], a categorical variable with three levels "1", "3", "5";
- HAND - the hand or hands used to perform the experiment, a categorical variable with three levels, "D" - dominant hand, "N" - non-dominant hand , "B" - both hands;

The goal is to study the influence of a circle size and hand/hands used to perform the experiment on a number of hits.

Mean values and variances

The following table provides the summary of the data set:

```
##  BLOCK    HITS_SUM     DIAMETER HAND
##  1:9   Min.   : 9.00   1:12     B:12
##  2:9   1st Qu.:15.75   3:12     D:12
##  3:9   Median :22.00   5:12     N:12
##  4:9   Mean   :21.75
##        3rd Qu.:26.00
##        Max.   :44.00
```

The following tables provide the mean values and variances for each variable.

BLOCK	1	2	3	4
mean value	19.00	19.67	22.00	26.33
variance	30.75	31.00	40.75	100.75

HAND	Both	Dominant	Non-Dominant
mean value	19.00	25.42	20.83
variance	29.82	82.45	38.70

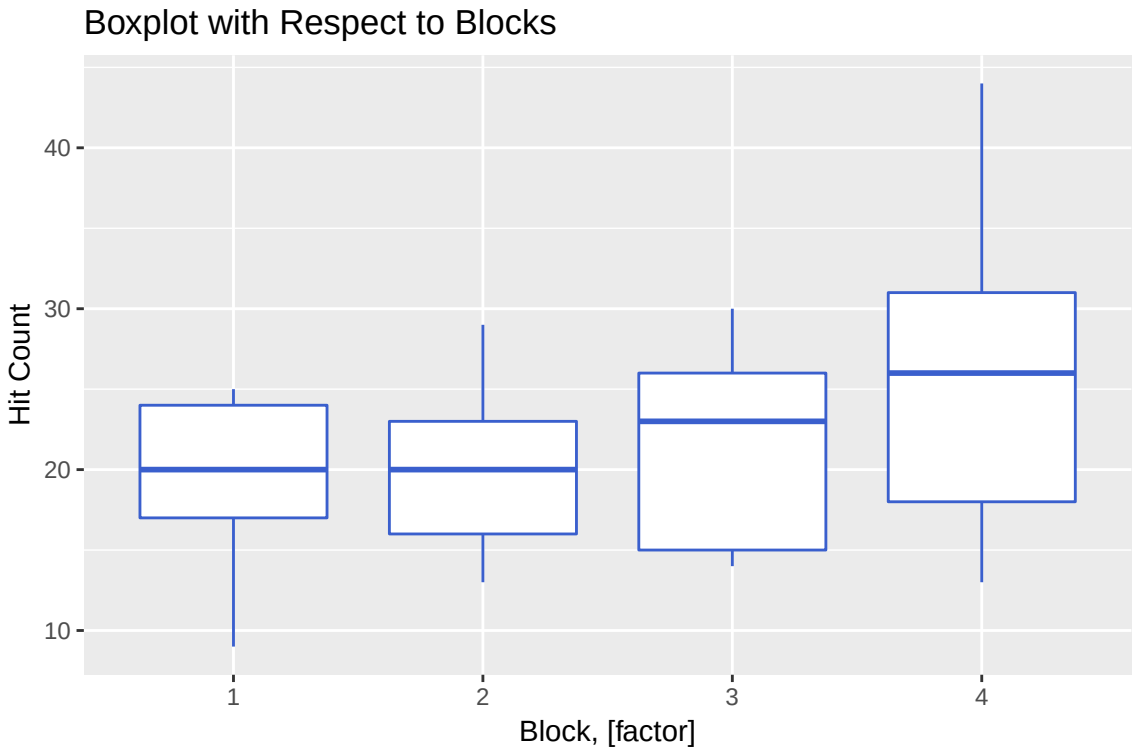
DIAMETER	1 cm	3 cm	5 cm
mean value	14.17	24.67	26.42
variance	5.42	20.97	52.63

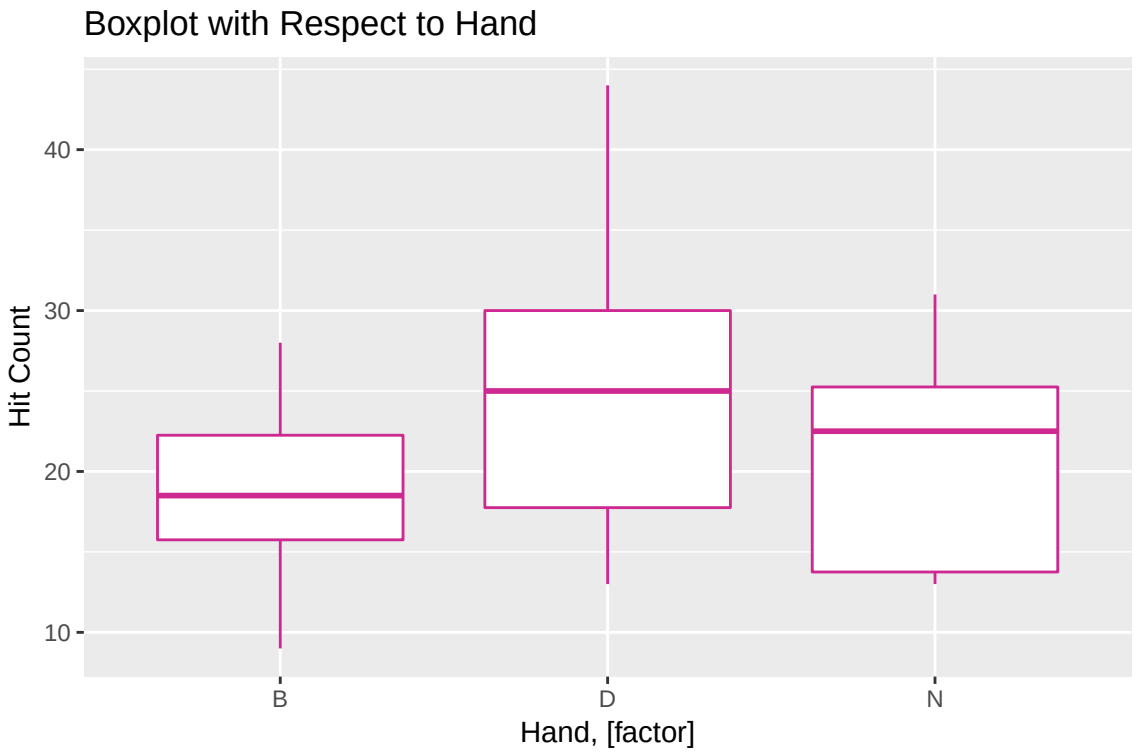
We can see that mean values for each of the blocks are slightly different. However, the 4th block shows an outstanding behavior. As a result, further investigation is needed. Regarding circle diameters, the data shows, that with bigger diameter the number of hits increases. Turning to mean values with respect to the hand, as expected, the number of hits made by the dominant hand is noticeably larger than that of the non-dominant and both hands. The variance shows the same behavior as that of mean values. Once again, the 4th block displays outlying performance.

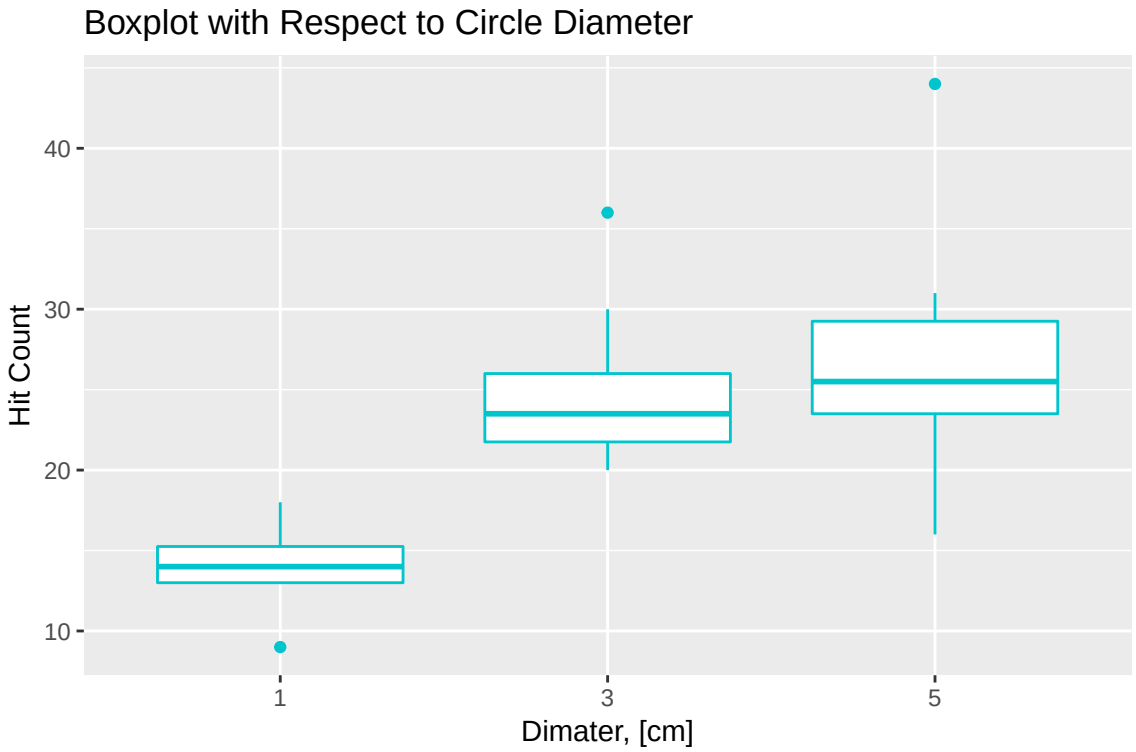
Data visualization

Let's visualize the dataset using boxplots and interaction plots.

Boxplots:

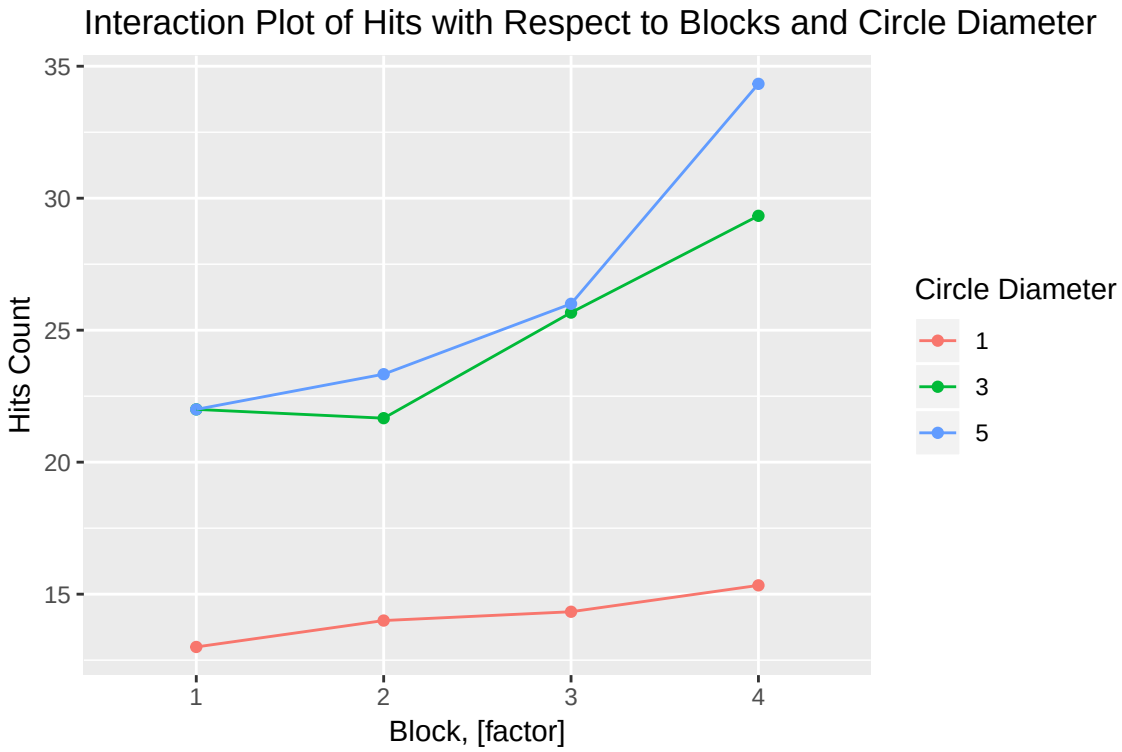


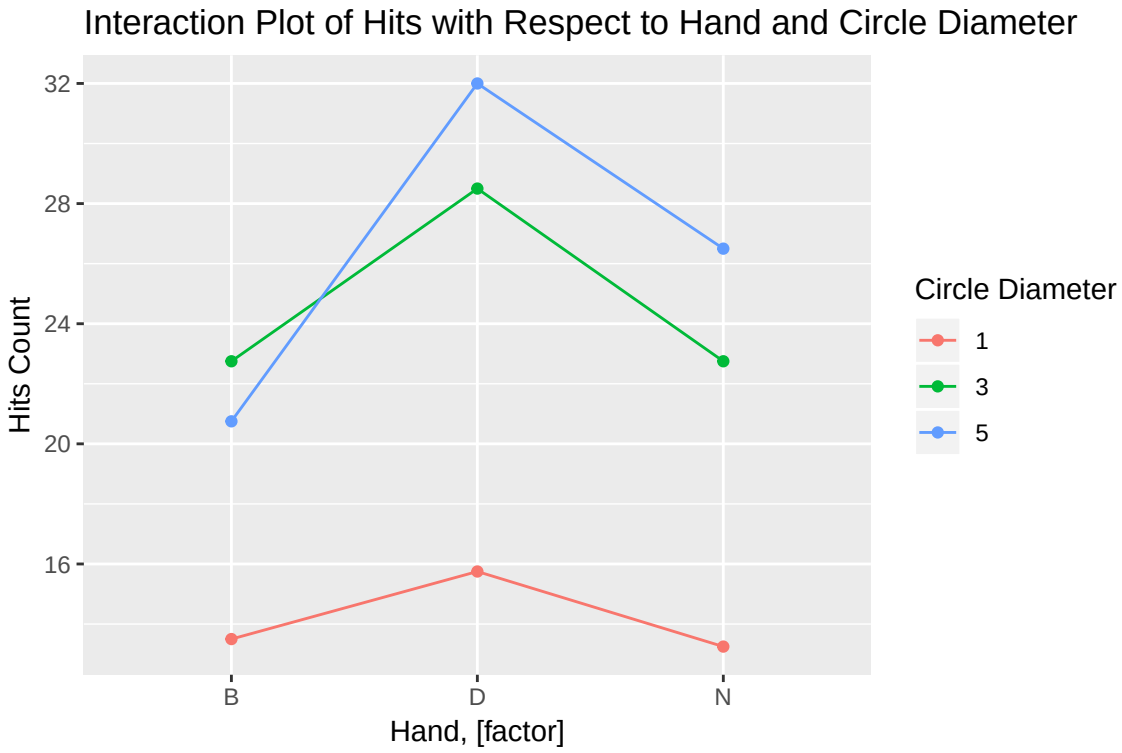


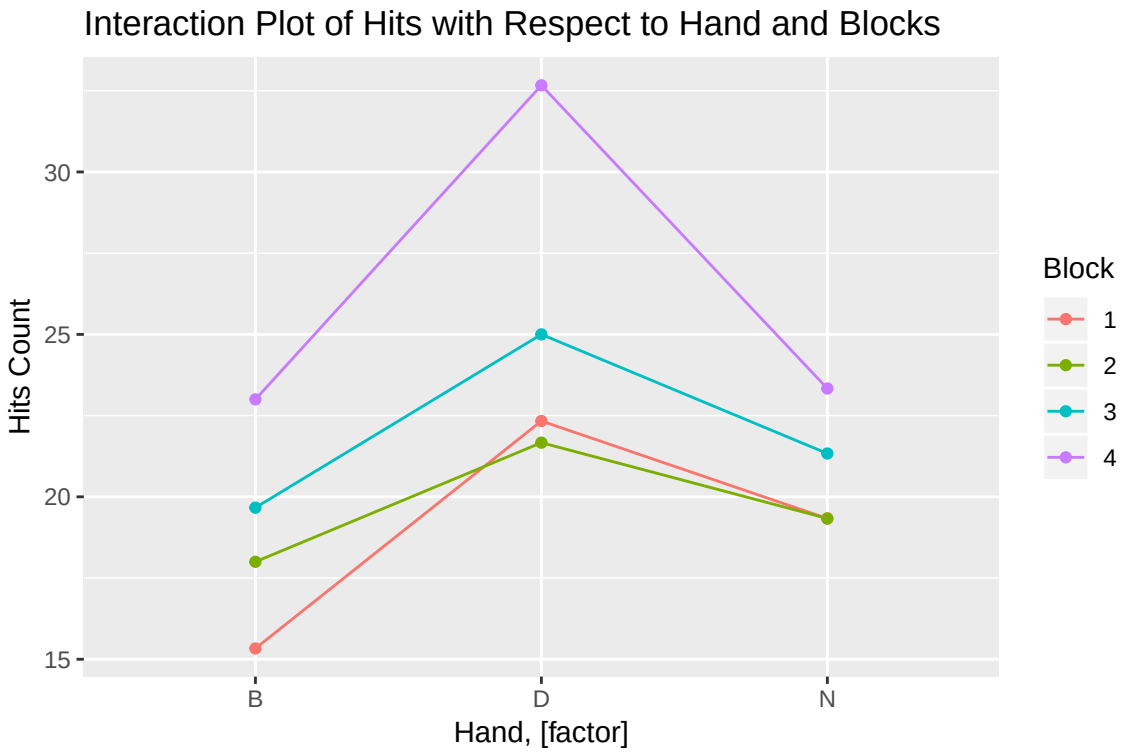


According to the boxplot visualization we can speculate, that mean values are significantly different for the "DIAMETER" variable. On the other hand, for 'HAND' and 'BLOCK' variables differences are not clear - relevant statistical tests have to be performed.

Interaction plots:







Interaction plots 1 and 3 display, that the 4th block (operator) is different from the rest. Others show similar circle hits count. This can possibly be caused by the effect of noise. Interection plot 2 displays the dependence of the circle hits count on the "HAND" and "DIAMETER" variables, e.g. hits count to the circle of diameter 5 cm for the dominant hand is the largest.

ANOVA

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## BLOCK        3  296.8    98.9   8.925 0.000261 ***
## HAND         2  262.2   131.1  11.827 0.000189 ***
## DIAMETER     2 1053.5   526.7  47.526 9.98e-10 ***
## Residuals   28  310.3    11.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

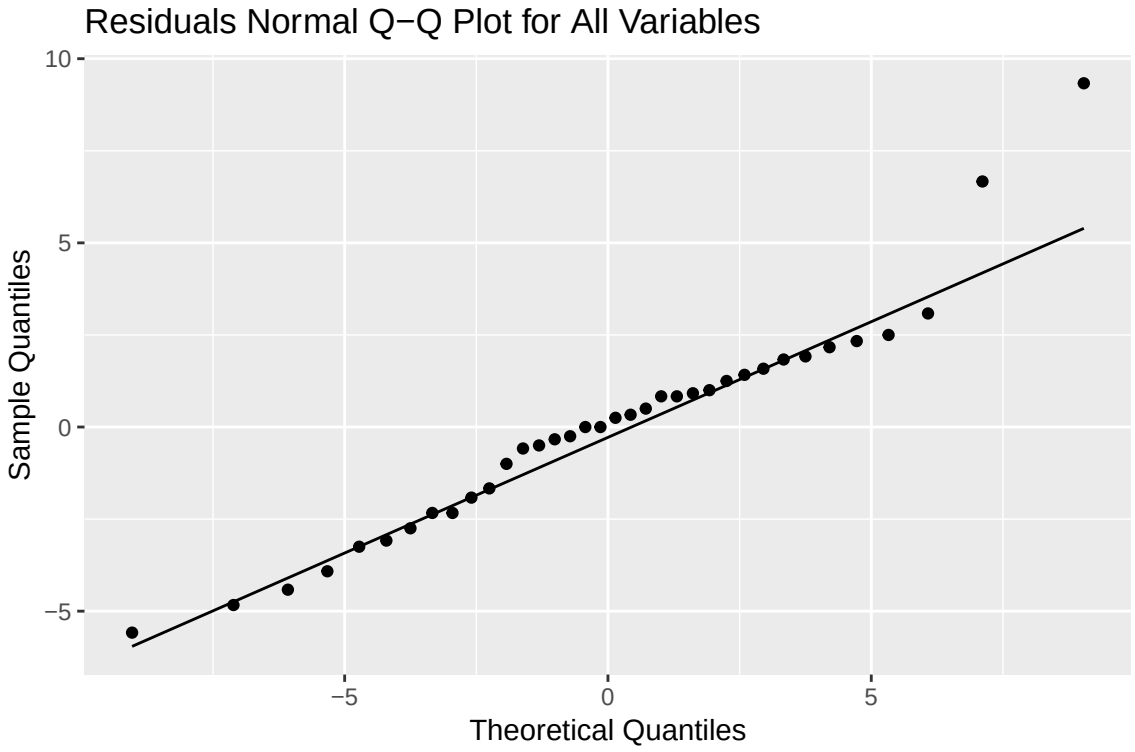
ANOVA has shown, that all variables are significant on the 95% significance level.

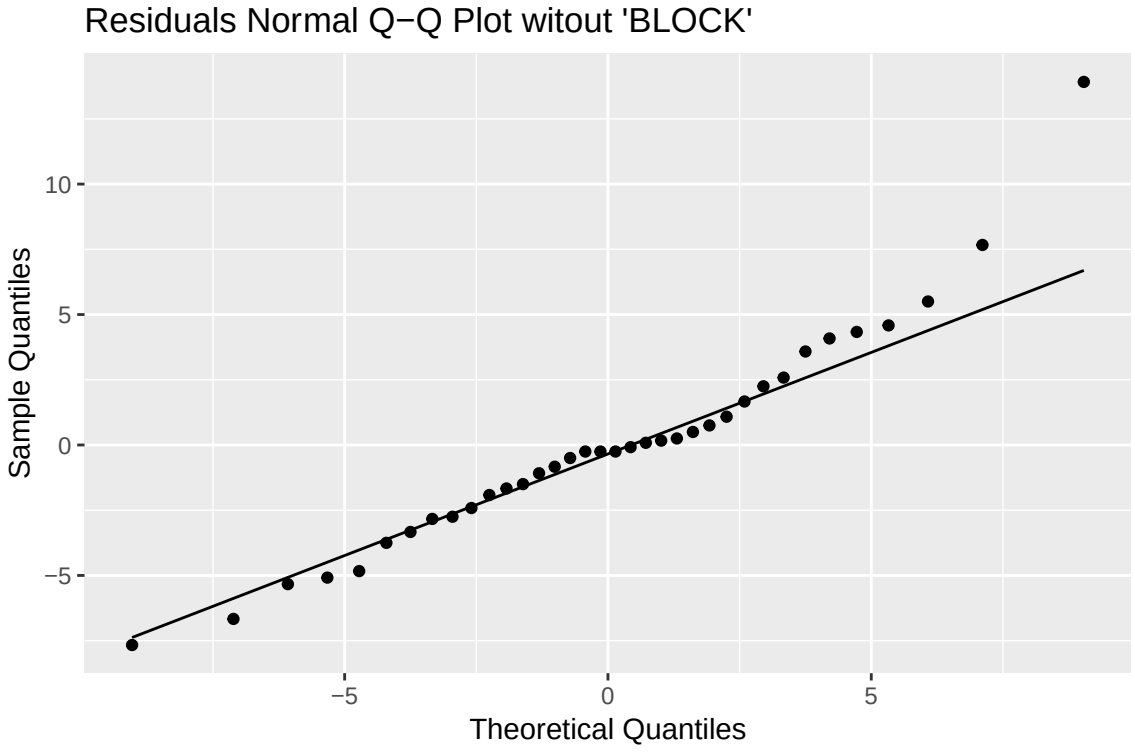
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## HAND         2  262.2   131.1   6.694  0.00383 **
## DIAMETER     2 1053.5   526.8  26.898 1.68e-07 ***
## Residuals   31  607.1    19.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variables "HAND" and "DIAMETER" are still significant even without dependence of the circle hits on the blocks (operators). That enables us to reject the hypothesis about the equality of mean values.

Residuals

Q-Q plot for residuals





Q-Q plots lines fit the data in an acceptable way. However, a few values display outlying behavior. Normality tests must be carried out. We perform Shapiro-Wilk test. The following is the result of the test.

Shapiro-Wilk normality test for residuals from the model with all variables and from the model without 'BLOCK' variable, respectively:

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_aov_all
## W = 0.94306, p-value = 0.06331
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_aov
## W = 0.94531, p-value = 0.07444
```

As p-values from the Shapiro-Wilk test are close to the set significance level (5%), we will also perform the Lilliefors test of normality.

Lilliefors normality test for residuals from the model with all variables and from the model without 'BLOCK' variable, respectively:

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuals_aov_all
## D = 0.11724, p-value = 0.2403
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuals_aov
## D = 0.12299, p-value = 0.1828
```

As a result of the test, we cannot reject the residuals normality hypothesis for both models.

Analysis of Factor Variables

Variable 'HAND'

Fisher's LSD-test:

```
##   hit_data$HITS_SUM      std  r      LCL      UCL Min Max   Q25  Q50   Q75
## B          19.00000 5.460603 12 17.02991 20.97009   9  28 15.75 18.5 22.25
## D          25.41667 9.080031 12 23.44657 27.38676  13  44 17.75 25.0 30.00
## N          20.83333 6.220689 12 18.86324 22.80343  13  31 13.75 22.5 25.25
```

```
##   hit_data$HITS_SUM groups
## D          25.41667      a
## N          20.83333      b
## B          19.00000      b
```

Tukey's HSD-test:

```
##          diff       lwr       upr        p adj
## D-B  6.416667  3.053714  9.779619 0.0001710998
## N-B  1.833333 -1.529619  5.196286 0.3808515200
## N-D -4.583333 -7.946286 -1.220381 0.0060219304
```

Both tests have confirmed, that the performance of the dominant hand is significantly different from other variants.

Variable 'BLOCK'

Fisher's LSD-test:

```
##   hit_data$HITS_SUM       std r      LCL      UCL Min Max Q25 Q50 Q75
## 1          19.00000  5.545268 9 16.72513 21.27487   9  25  17  20  24
## 2          19.66667  5.567764 9 17.39180 21.94154  13  29  16  20  23
## 3          22.00000  6.383573 9 19.72513 24.27487  14  30  15  23  26
## 4          26.33333 10.037430 9 24.05846 28.60820  13  44  18  26  31
```

```
##   hit_data$HITS_SUM groups
## 4          26.33333      a
## 3          22.00000      b
## 2          19.66667      b
## 1          19.00000      b
```

Tukey's HSD-test:

```
##          diff         lwr       upr        p adj
## 2-1 0.6666667 -3.61823787  4.951571 0.9737562956
## 3-1 3.0000000 -1.28490454  7.284905 0.2462304955
## 4-1 7.3333333  3.04842879 11.618238 0.0003753571
## 3-2 2.3333333 -1.95157121  6.618238 0.4585614479
## 4-2 6.6666667  2.38176213 10.951571 0.0011686578
## 4-3 4.3333333  0.04842879  8.618238 0.0467076034
```

Once again, we observe significant difference between the 4th block (operator) and 3 other blocks. An interesting observation is that the 3rd block is on the edge of being significantly similar to the 4th one.

Variable 'DIAMETER'

Fisher's LSD-test:

```
##   hit_data$HITS_SUM      std  r      LCL      UCL Min Max   Q25  Q50   Q75
## 1          14.16667 2.329000 12 12.19657 16.13676   9  18 13.00 14.0 15.25
## 3          24.66667 4.579268 12 22.69657 26.63676  20  36 21.75 23.5 26.00
## 5          26.41667 7.254570 12 24.44657 28.38676  16  44 23.50 25.5 29.25
```

```
##   hit_data$HITS_SUM groups
## 5          26.41667      a
## 3          24.66667      a
## 1          14.16667      b
```

Tukey's HSD-test:

```
##      diff       lwr       upr        p adj
## 3-1 10.50  7.137047 13.862953 6.047549e-08
## 5-1 12.25  8.887047 15.612953 2.686814e-09
## 5-3  1.75 -1.612953  5.112953 4.137523e-01
```

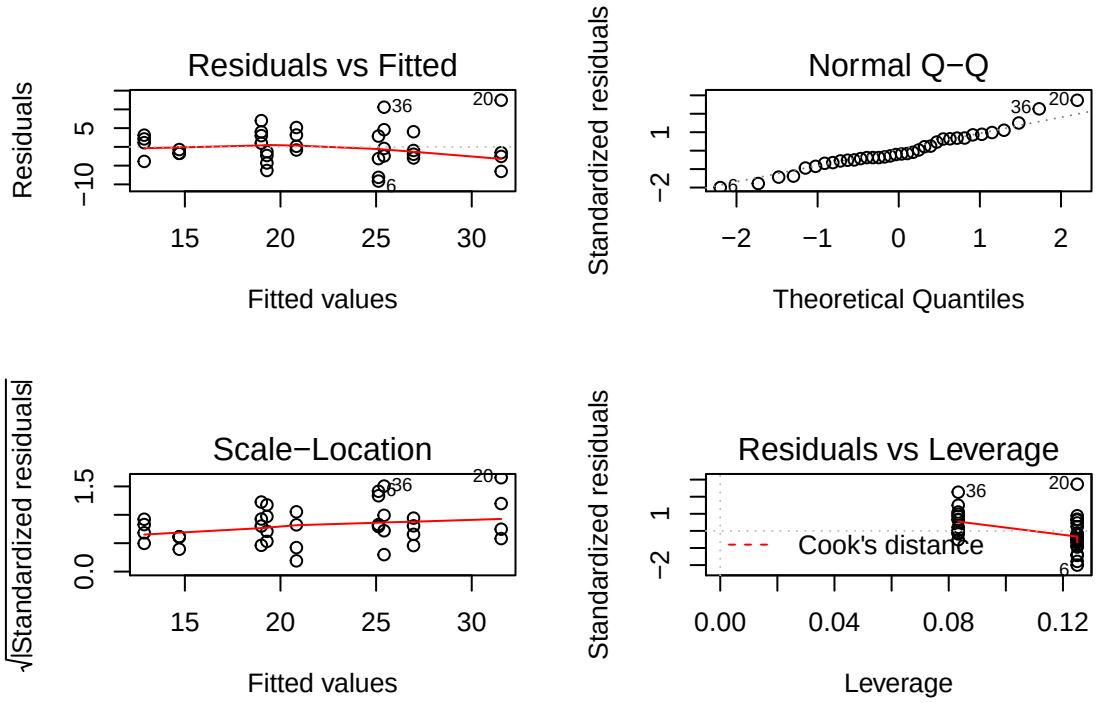
Tukey's HSD test and Fisher's LSD test indicate, that circles with diameters 3 cm and 5 cm are significantly similar. On the other hand, the circle with diameter of 1 cm is significantly different from two other ones.

Linear Regression

We fit a linear model without intercept, where we consider the variable "DIAMETER" and the variable "HAND":

$$\mathbb{E}(HITS\_SUM|DIAMETER, HAND) = \beta_1 DIAMETER + \beta_2 HAND.$$

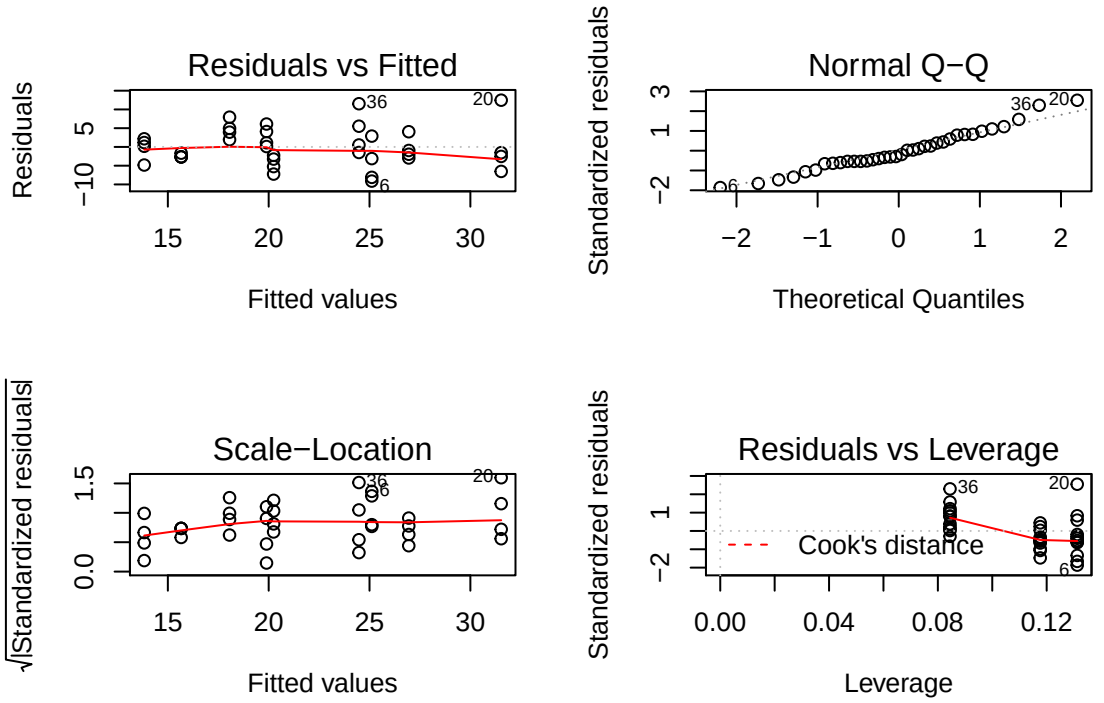
```
##
## Call:
## lm(formula = HITS_SUM ~ -1 + DIAMETER + HAND, data = lm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1250 -2.4479 -0.8958  3.0312 12.4583
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## DIAMETER   6.1250     0.9949   6.156 6.91e-07 ***
## HANDB      6.7500     2.4370   2.770  0.00926 **
## HANDD     13.1667     2.4370   5.403 6.15e-06 ***
## HANDN      8.5833     2.4370   3.522  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.874 on 32 degrees of freedom
## Multiple R-squared:  0.9599, Adjusted R-squared:  0.9549
## F-statistic: 191.5 on 4 and 32 DF,  p-value: < 2.2e-16
```



We fit another linear model, where we consider the variable "DIAMETER" set to the power of 2 and the variable "HAND":

$$\mathbb{E}(HITS\_SUM|DIAMETER, HAND) = \beta_1(DIAMETER^2) + \beta_2 HAND.$$

```
##
## Call:
## lm(formula = HITS_SUM ~ -1 + I(DIAMETER^2) + HAND, data = lm_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.113 -2.732 -1.211  3.150 12.470
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## I(DIAMETER^2)   1.4107     0.2649   5.325 7.72e-06 ***
## HANDB          12.4167     1.9548   6.352 3.94e-07 ***
## HANDD          18.8333     1.9548   9.634 5.60e-11 ***
## HANDN          14.2500     1.9548   7.290 2.76e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.245 on 32 degrees of freedom
## Multiple R-squared:  0.9535, Adjusted R-squared:  0.9477
## F-statistic: 164.2 on 4 and 32 DF,  p-value: < 2.2e-16
```



Normality of residuals

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_circle_1$residuals
## W = 0.96522, p-value = 0.3096
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_circle_2$residuals
## W = 0.96775, p-value = 0.3674
```

As Q-Q plots and Shapiro-Wilk test indicate, general assumptions for performing the linear regression task are met. According to the R-squared statistic, the model with the circle diameter set to the power of 2 explains the hit data slightly worse. However, the difference is negligible. As a result, we choose the first model.