NEX: Homework Assignment 02

Belov, Neoral, Sahan, Shulga

In this assigment we use data set that was collected from an experiment performed by FNSPE CTU students. The goal of this experiment was to investigate what factors can have possible effect on the students precision while throwing filled rubber balloons. To collect the data, four students were throwing ballons aiming to hit a line. The following factors were taken into consideration: the distance from a testee to the line, a balloon filler, a baloon mass, a hand used to throw a balloon, opened or closed eyes of a testee and testee stance. The dataset consists of 80 observations of 6 categorical variables and responce variable MEASUREMENT. All categorical variables have two levels.

- MASS - a mass of a balloon, “-1” - 50g, “1” - 100g;
- DISTANCE - a distance from the line, “-1” - 3 m, “1” - 5 m;
- FILLING - balloon filling, “-1” - grain, “1” - flour;
- HAND - hand used to throw a balloon, “-1” - dominant, “1” - non-dominant;
- VISION - “-1” - eyes of a testee were opened, “1” - eyes of a testee were closed;
- STANCE - “-1” - a testee stood on two legs, “1” - a testee stood on one leg;
- MEASUREMENT - a distance from a balloon to the line, measured from the furthest part of a ballon from the line, in mm.

Last 16 observations were made with central poits for MASS and DISTANCE, i.e with balloons of the mass 75 g and the distance from the line equal to 4 m.

Experiment design

Let’s denote the first factor MASS by A, the second factor DISTANCE by B, FILLING by C, HAND by D, VISION by E, STANCE by F. Each student performed one of four alternate fractions of one-quarter fraction of the 2^6 Design. For the first quarter we chose I=ABCE and I=BCDF as the design generators. The complete defining relation in this case is

$$I=ABCE=BCDF=ADEF.$$

Thus, each effect has three aliases. For example for A we obtain

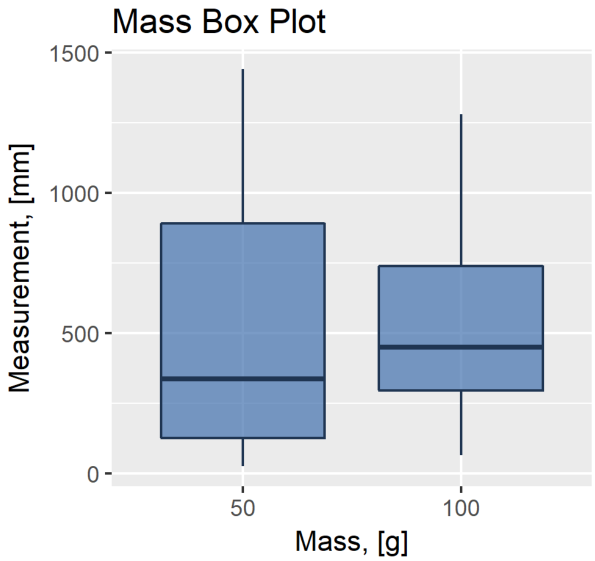
$$A=BCE=ABCDF=DEF$$

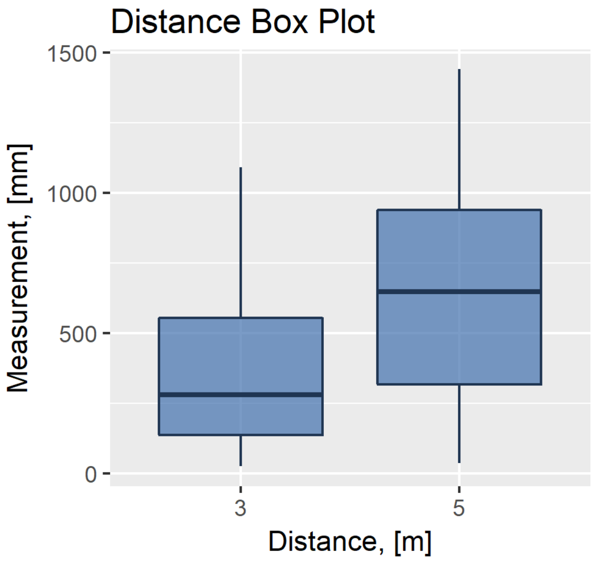
i.e. main effects are aliased by three- and five-factor interactions, two-factor interactions are aliased with each other and with interactions of higher order. Hence our design is a design of resolution IV. There are three alternate fractions of this design, these are fractions with following generating relations

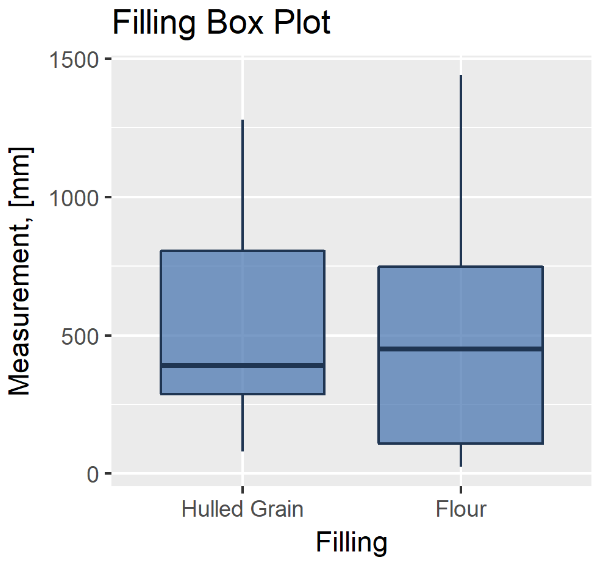
$$I=ABCE \text{ and } I=-BCDF, I=-ABCE \text{ and } I=BCDF, I=-ABCE \text{ and } I=-BCDF.$$

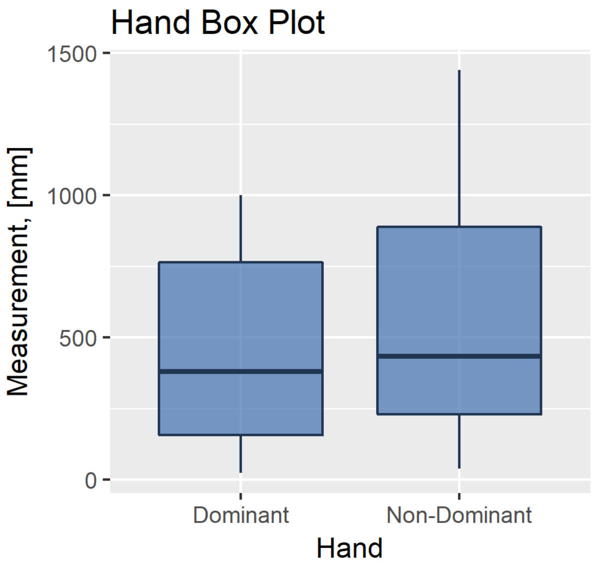
Basic visual analysis

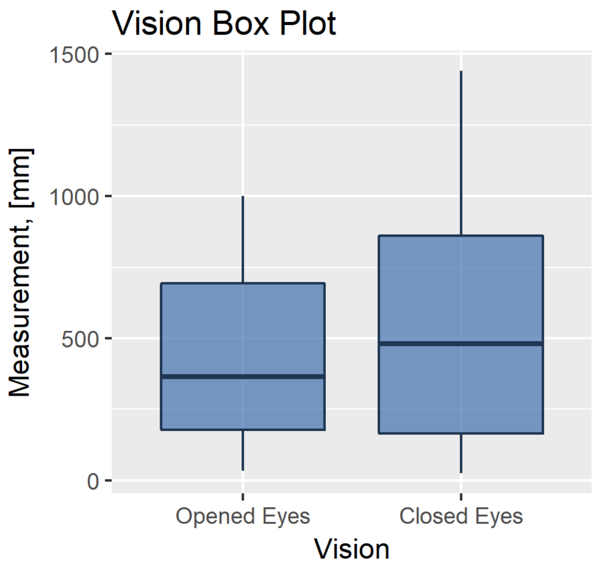
The Figure on page 2 provides the box plots for all factor variables. A noticeable observation is that only the box plot for the variable DISTANCE indicates remarkable differences in MEASUREMENT with respect to the factor values. On the other hand other variables do not display any visible differences.

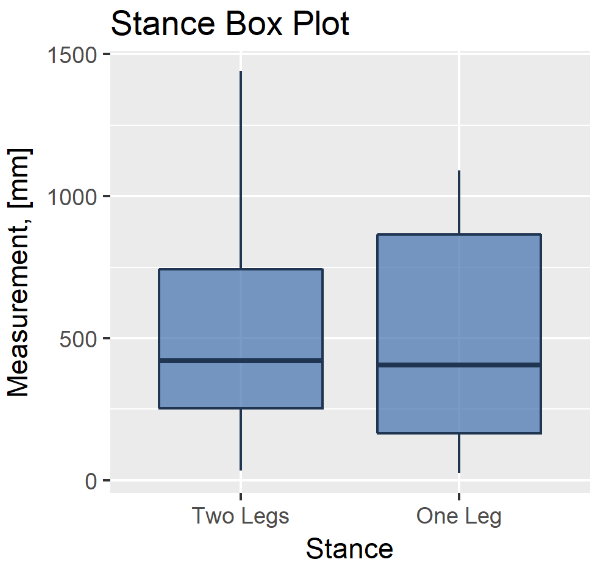

Mass Box Plot


Distance Box Plot


Filling Box Plot


Hand Box Plot


Vision Box Plot


Stance Box Plot

Main effects

A One-Way ANOVA tests if the mean value of the response variable differs among two or more levels of a factor. The following table provides a result of the test. The mean value of the response is statistically significantly different only for different distances and there is no statistically significant difference for other factors.

```
##             Df  Sum Sq Mean Sq F value  Pr(>F)   
## mass         1    5625    5625   0.047 0.82926   
## distance     1 1102500 1102500   9.199 0.00364 **
## filling      1   78400   78400   0.654 0.42200   
## hand         1   51189   51189   0.427 0.51604   
## vision       1  138756  138756   1.158 0.28647   
## stance       1    7439    7439   0.062 0.80415   
## Residuals   57 6831491  119851                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

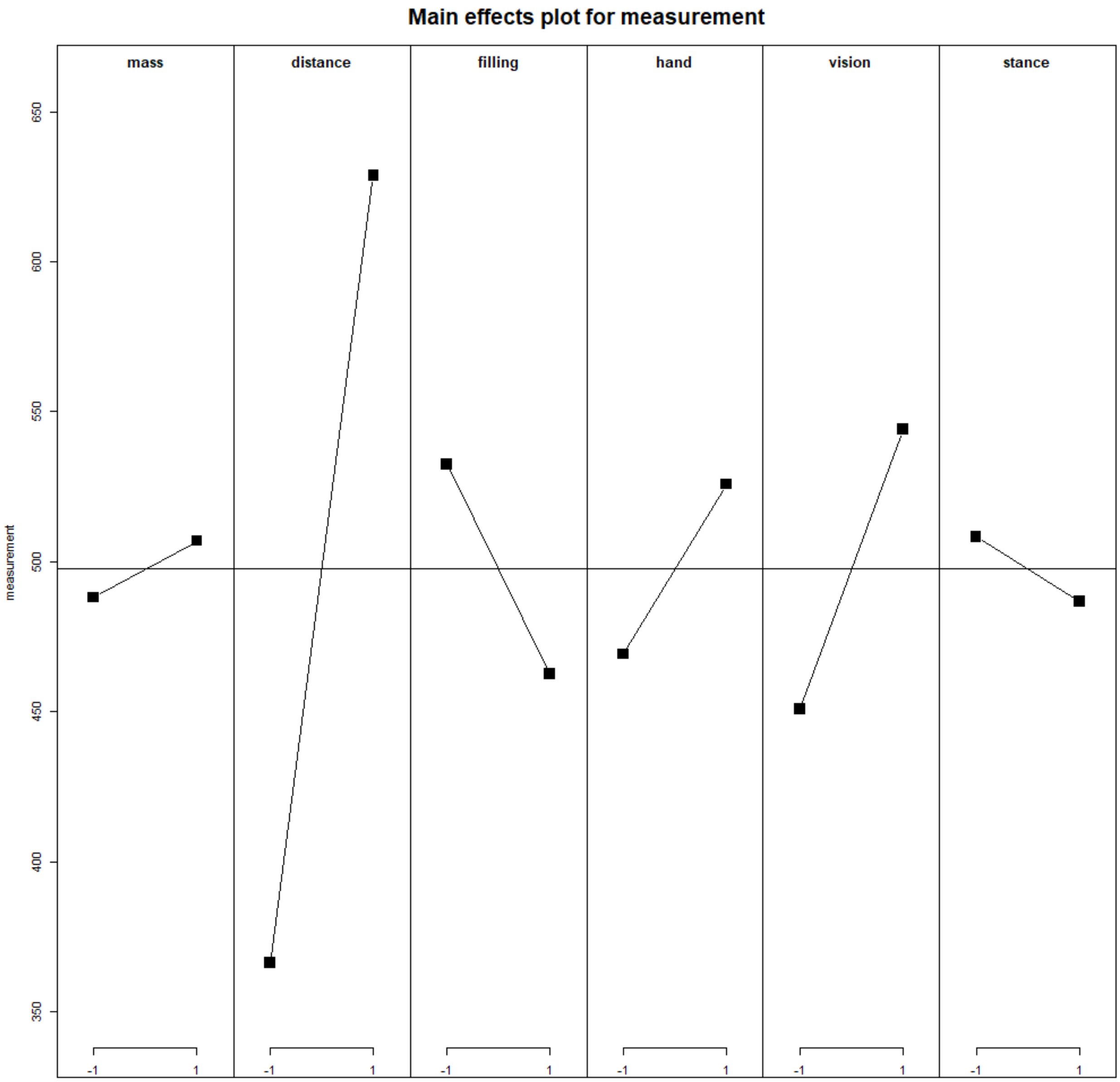
The average effect of a factor is defined as the change in response produced by a change in the level of that factor averaged over the levels of the other factors. Let's examine magnitudes and directions of the main effects to determine which factors are likely to be important. The main effects of all factors are plotted in Figure on page 4. The effects of MASS, DISTANCE, HAND and VISION are positive, this suggests that an increase of this factors from the low level to the high level increases the value of the response. The effects of FILLING and STANCE are negative, this suggests that an increase of this factors from the low level to the high level decreases the value of the response variable MEASURMENT. The main effect of all factors appear to be small relative to the main effect of DISTANCE.

Interactions

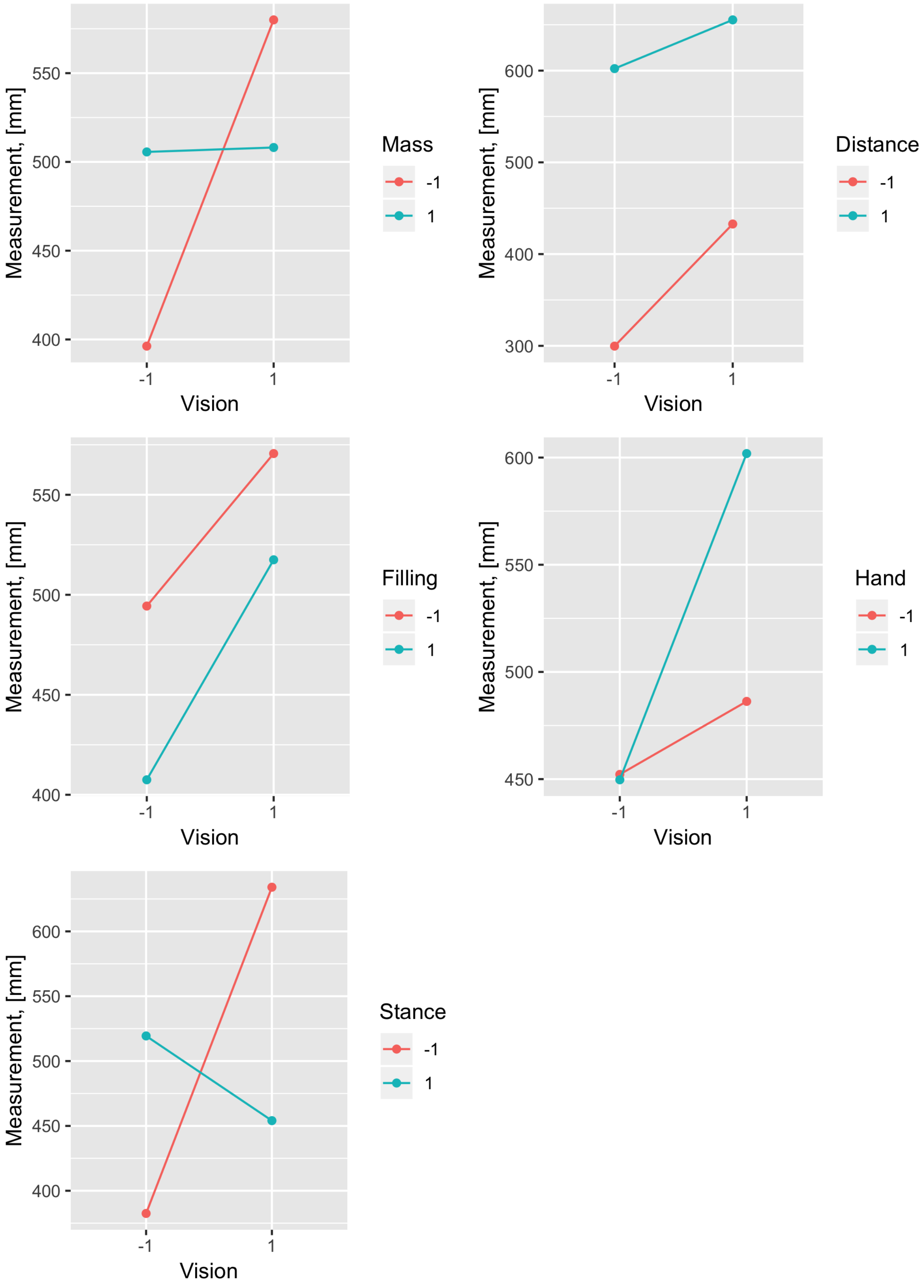
The two-factor interactions are plotted in Figures on pages 5-10. Looking at interaction plots for VISION we can note that there is almoust no effect of VISION when MASS is at the high level, while there is lange effect of VISION when MASS is at the low level; the effect of VISION is very large when HAND is at the high level ad very small when HAND is on the low level; the effect of VISION at the high level of STANCE is negative while it is positive at the low level of STANCE. If we look at the interaction plots for STANCE we can see that the effect of STANCE is bigger at the high level of MASS than at the low lever of MASS; the effect of STANCE is positive at the low level of HAND and negative at the high level of HAND and it shows same behaviour at levels of VISION. It is important to notice, that the effect of DISTANCE is always positive at all levels of other factors.

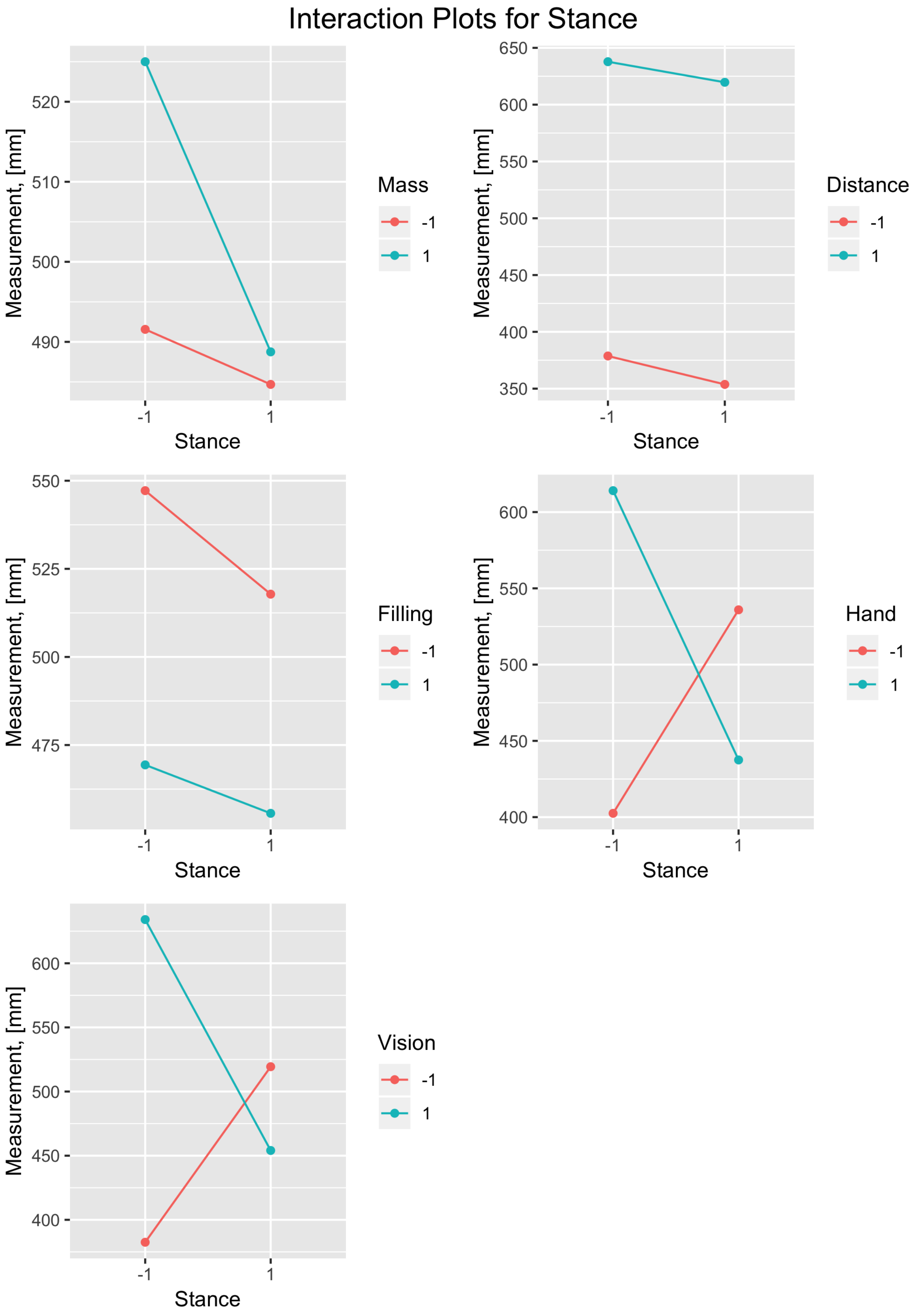
The normal probability plot of the effects is shown in Figure on page 11. All of the effects that lie along the line are negligible, the large effects are far from the line. The important effects that emerge from this analysis are those that are signed, for instance the effect of the distance, the effect of the interaction of mass, distance, filling, hand and stance. These effects are significant at the 5% significance level.

We performed the ANOVA test for all possible interactions. The result of the test was that none of the interactions was significant.

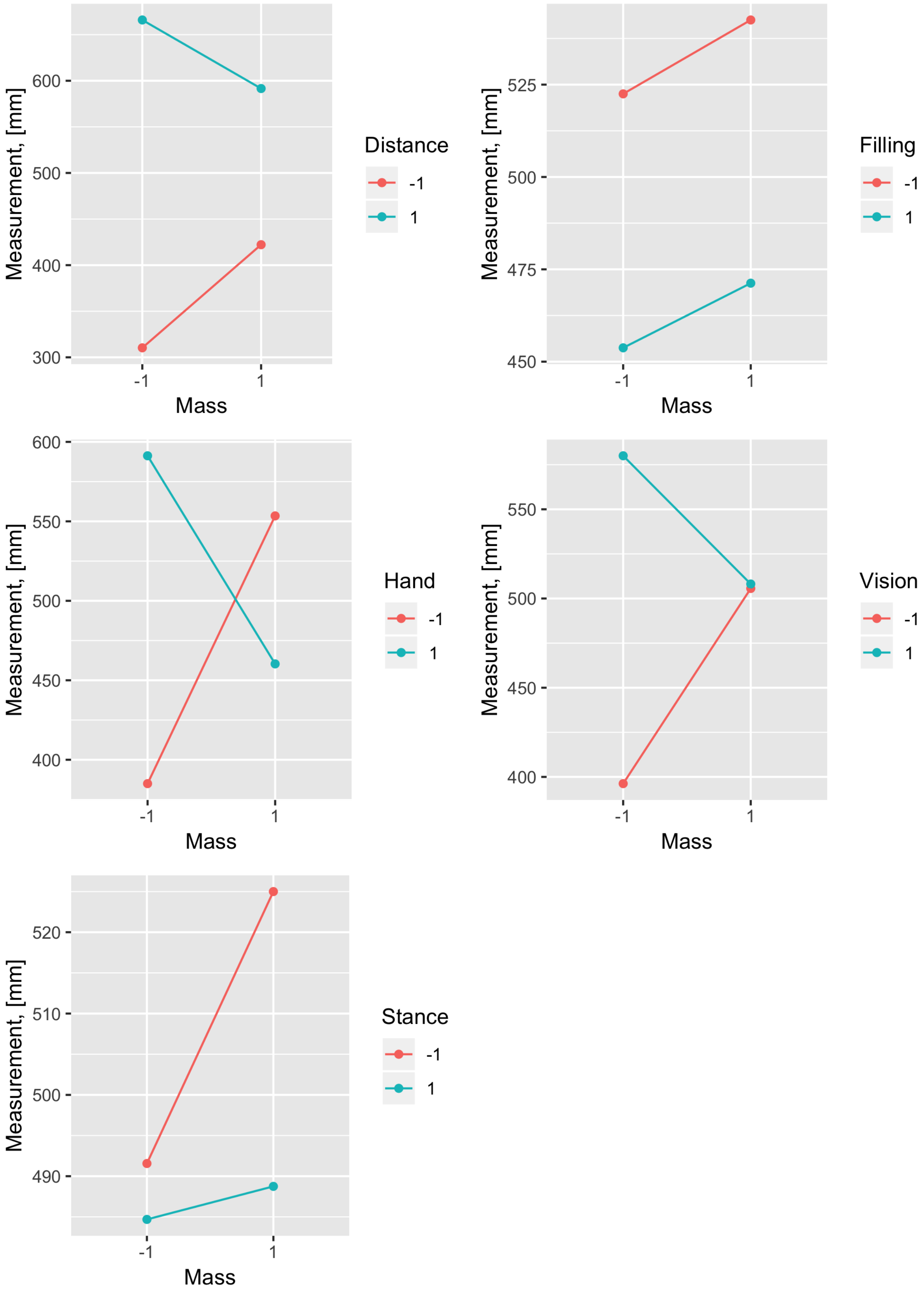

Main effects plot for measurement
mass
distance
filling
hand
vision
stance
measurement
650
600
550
500
450
400
350
-1
1
-1
1
-1
1
-1
1
-1
1
-1
1

Interaction Plots for Vision

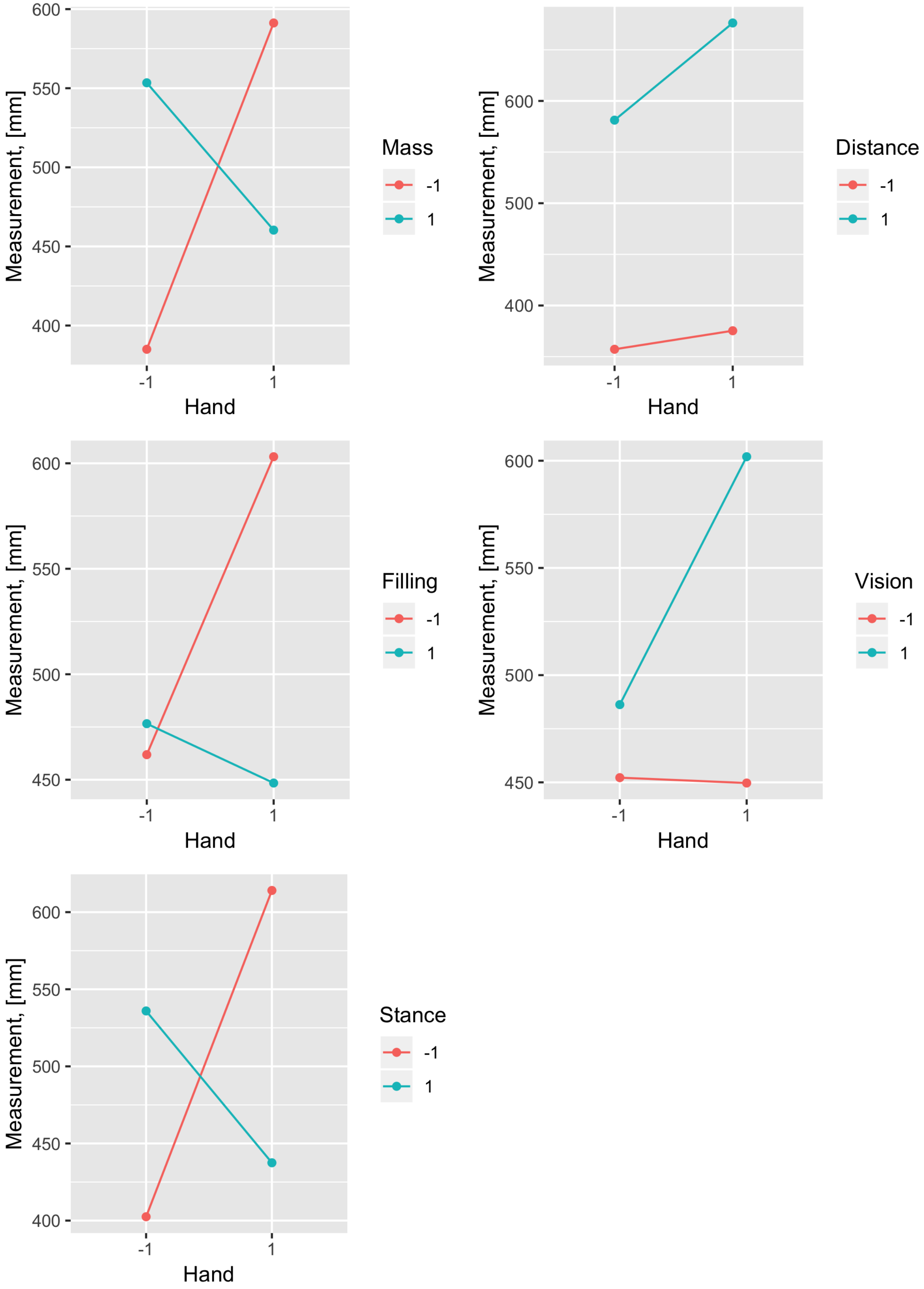



Interaction Plots for Stance
Measurement, [mm]
520
510
500
490
-1
1
Stance
Mass
-1
1
Measurement, [mm]
650
600
550
500
450
400
350
-1
1
Stance
Distance
-1
1
Measurement, [mm]
550
525
500
475
-1
1
Stance
Filling
-1
1
Measurement, [mm]
600
550
500
450
400
-1
1
Stance
Hand
-1
1
Measurement, [mm]
600
550
500
450
400
-1
1
Stance
Vision
-1
1

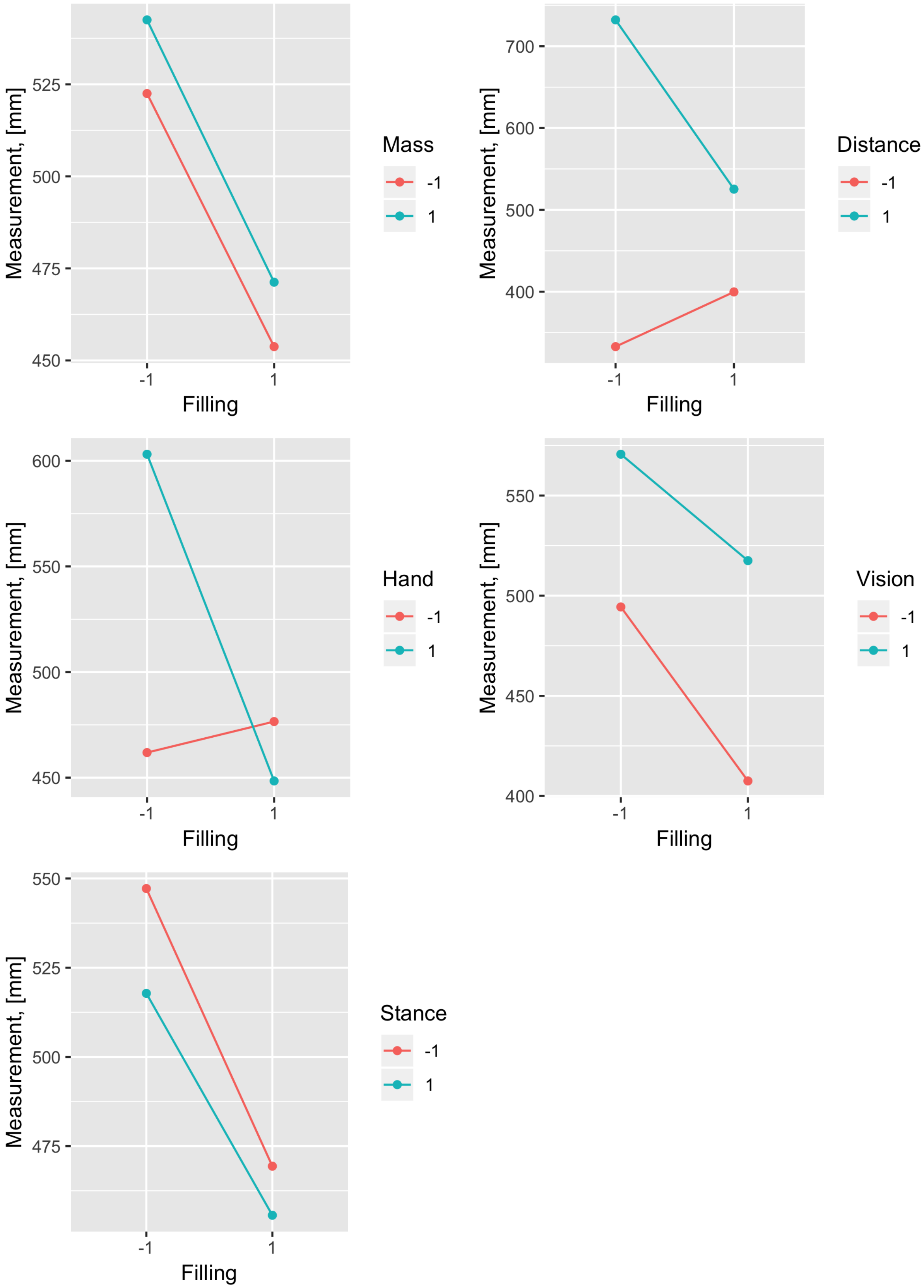
Interaction Plots for Mass



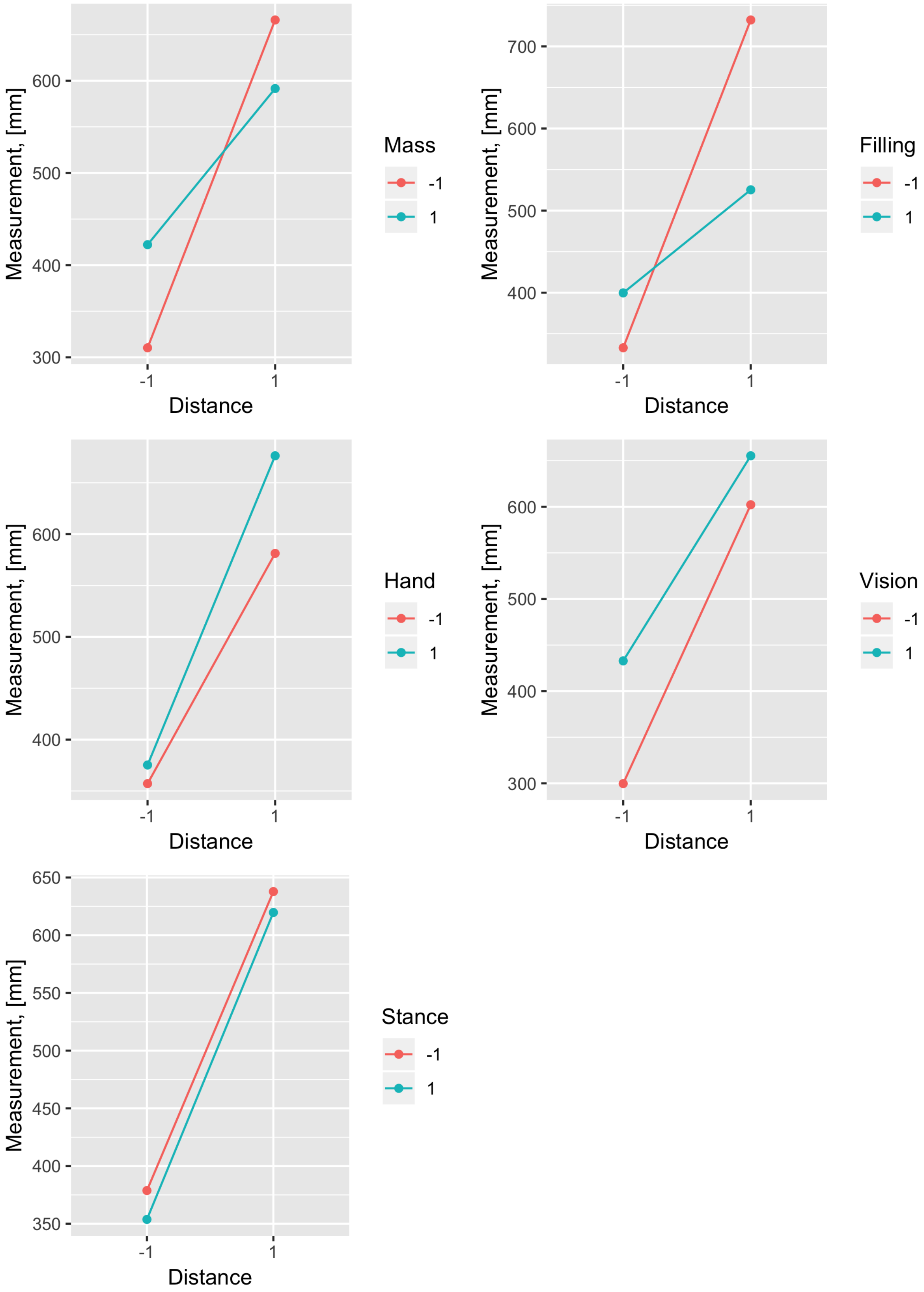
Interaction Plots for Hand

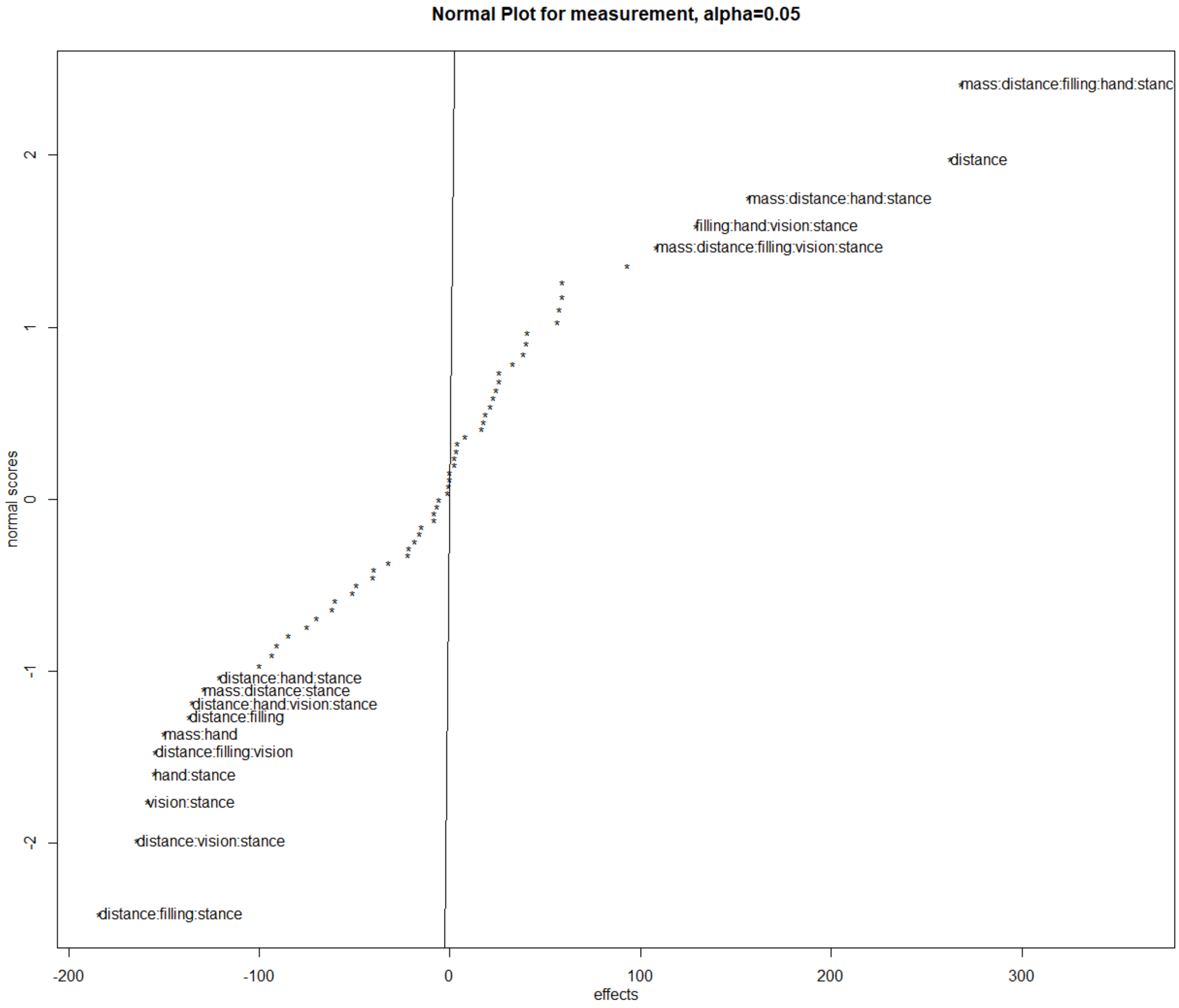


Interaction Plots for Filling

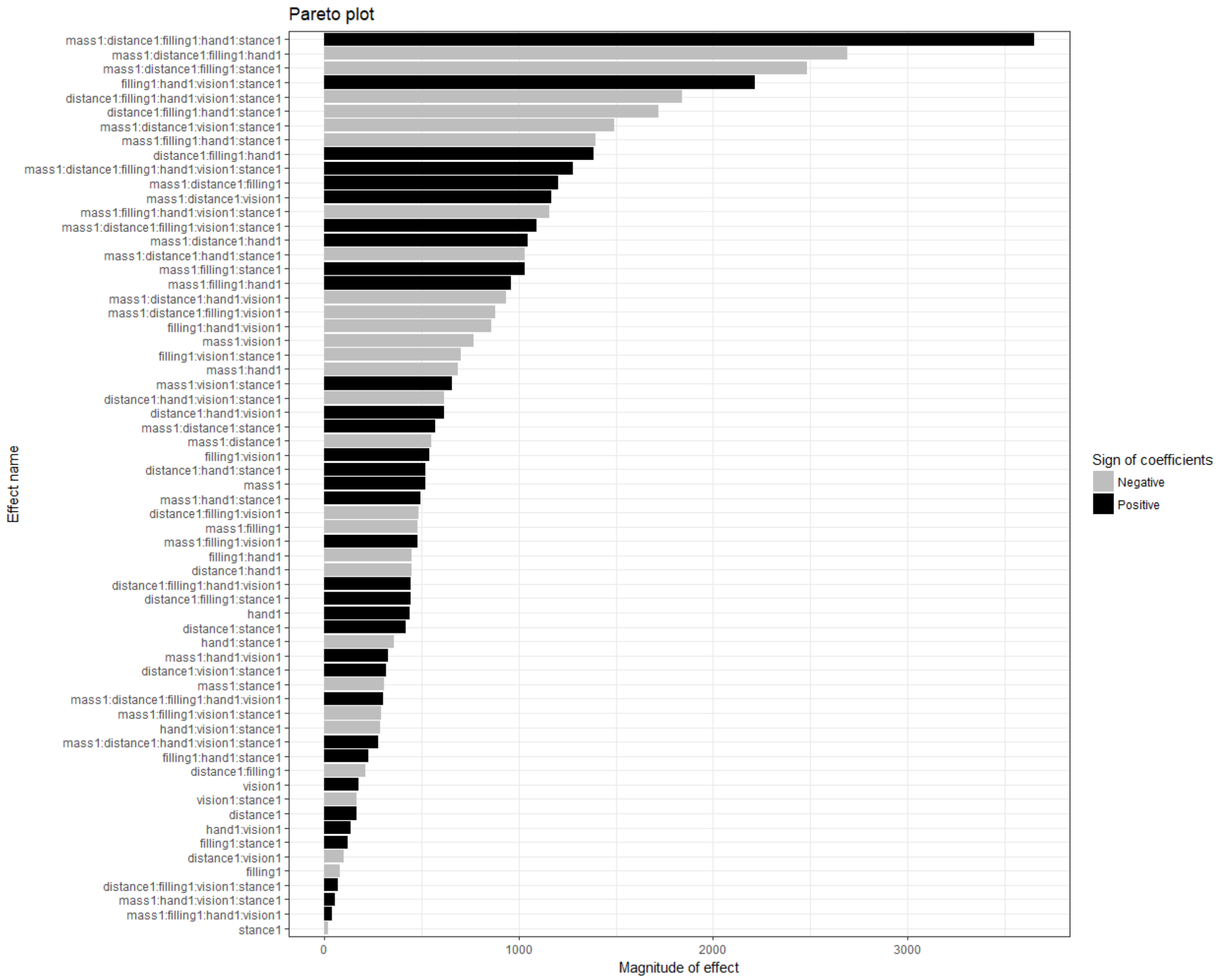


Interaction Plots for Distance





The following Pareto plot provides magnitudes of effects with their directions.



ANOVA

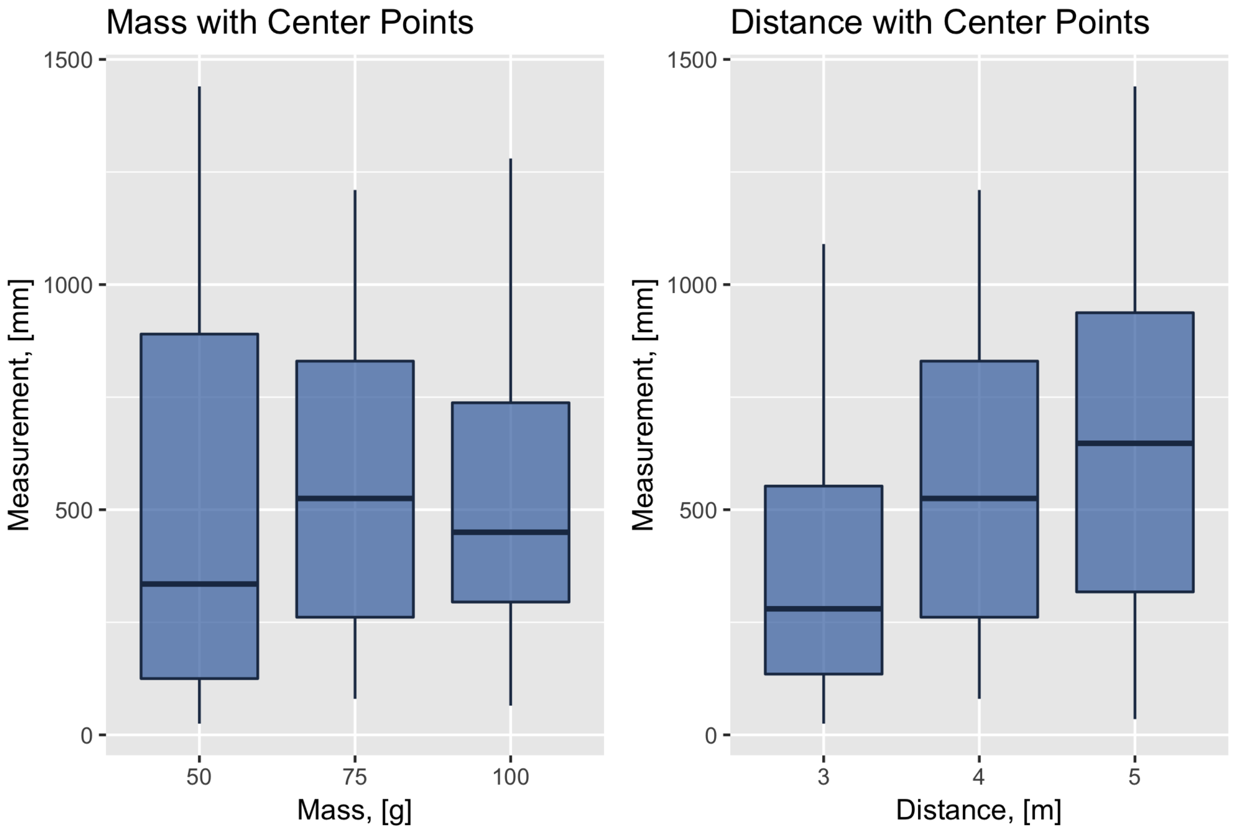
Let's create and investigate model with different interaction that were chosen with respect to the results of the analysis above. First we consider a linear model with all interactions that are significant according to normal plot on page 11. We perform the ANOVA test. The result of the test is that only DISTANCE is significant. Also the interaction DISTANCE:STANCE:VISION is close to be significant, thus we include the interaction in the next model.

The following table provides ANOVA results for final model.

```
##                         Df   Sum Sq Mean Sq F value Pr(>F)    
## distance                 2 16942900 8471450  77.661 <2e-16 ***
## distance:stance:vision   6  1004287  167381   1.534  0.184    
## Residuals               56  6108612  109082                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Center points

The following picture is box plots of variables with center points. The box plot “Distance with Center Points” speaks in favour of a linear dependance between DISTANCE and MEASURMENT.



We fit a linear model of measurement that depends on mass and distance without intercept, following is the summary for the model and results of ANOVA test.

```
##
## Call:
## lm.default(formula = measurement ~ mass + distance - 1, data = data_all)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -584.4 -284.0  -25.0  247.5  820.6
##
## Coefficients: (1 not defined because of singularities)
##           Estimate Std. Error t value Pr(>|t|)
## mass-1      356.88      74.13   4.814 7.36e-06 ***
## mass0       563.44      85.60   6.582 5.32e-09 ***
## mass1       375.62      74.13   5.067 2.76e-06 ***
## distance0       NA         NA      NA       NA
## distance1   262.50      85.60   3.067    0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 342.4 on 76 degrees of freedom
## Multiple R-squared:  0.712,  Adjusted R-squared:  0.6969
## F-statistic: 46.97 on 4 and 76 DF,  p-value: < 2.2e-16
```

```
##           Df   Sum Sq Mean Sq F value Pr(>F)
## mass       3 20925414 6975138  59.497 <2e-16 ***
## distance   1  1102500 1102500   9.404  0.003 **
## Residuals 76  8909811  117234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear regression

We fit another linear model

$$\mathbb{E}(MEASURMENT|MASS, DISTANCE) = \beta_1 MASS + \beta_2 DISTANCE,$$

where we treat MASS and DISTANCE as numerical variables.

```
##
## Call:
## lm.default(formula = measurement ~ mass + distance - 1, data = data_all_num)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -597.2 -276.8  -39.5  257.8  807.8
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## mass       0.1776     1.3682   0.130    0.897
## distance 124.6705    26.1252   4.772 8.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 339.1 on 78 degrees of freedom
## Multiple R-squared:  0.7101, Adjusted R-squared:  0.7026
## F-statistic: 95.51 on 2 and 78 DF,  p-value: < 2.2e-16
```

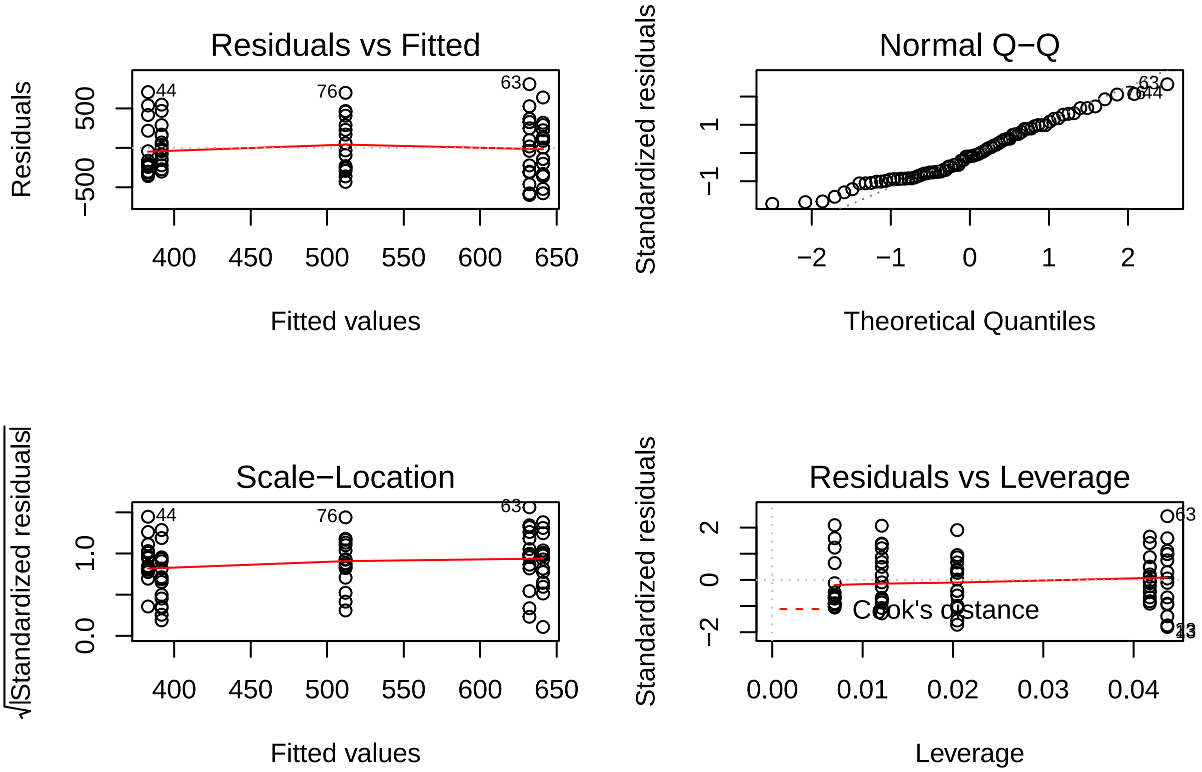
Let's run Lilliefors test and Shapiro-Wilk test to test normality of residuals.

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuals(final.lm_num)
## D = 0.11089, p-value = 0.01644
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(final.lm_num)
## W = 0.96523, p-value = 0.02833
```

Based on the p-values we reject the null hypothesis about the normality of the residuals.

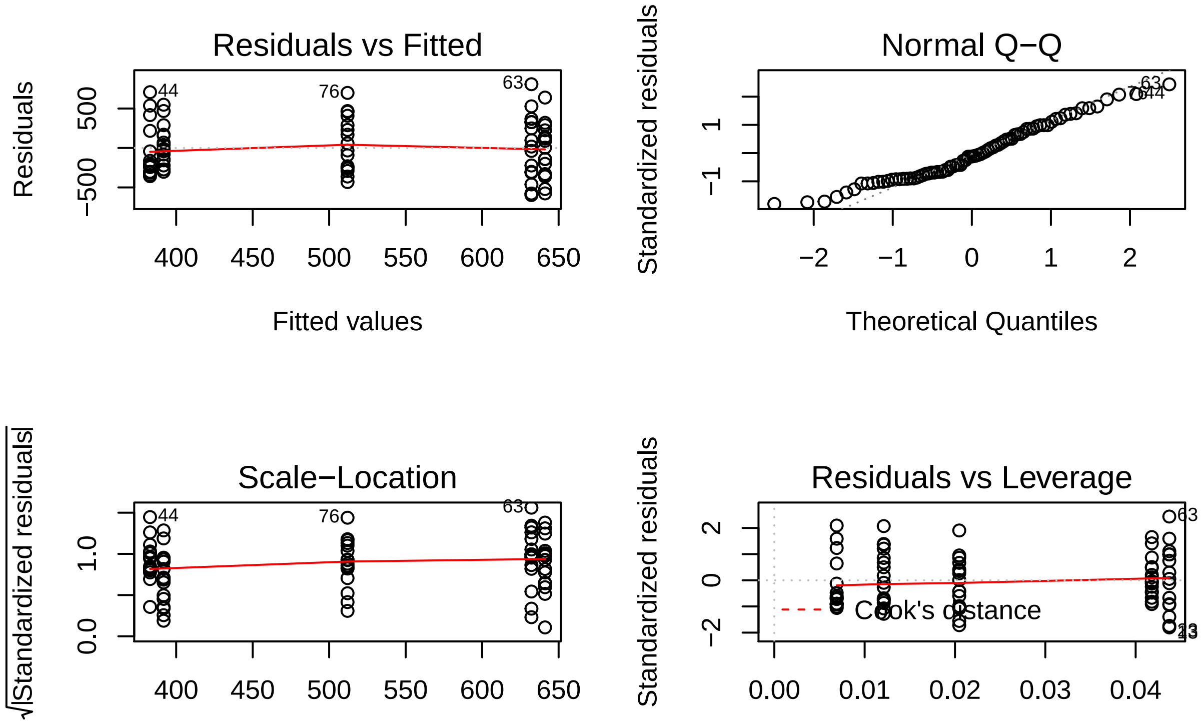
The following Figure provides residual plots for the model. Based on those plots, we can talk about the adequacy of the model.



We run the Breusch-Pagan test to detect heteroskedasticity. The following is the result of the test. Based on the p-value we reject the null hypothesis of homoskedasticity and assume heteroskedasticity.

```
##
##  studentized Breusch-Pagan test
##
## data:  final.lm_num
## BP = 5.322, df = 1, p-value = 0.02106
```

We considere the Box-Cox transformation to possibly improve the model. The following Figure provides residual plots for the transformed model.



The summary of the model with transformed response variable:

```
##
## Call:
## lm.default(formula = measurement_bc ~ mass + distance - 1, data = data_all_num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.437  -5.848   0.718   7.218  20.036
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## mass      0.08529    0.03647   2.339   0.0219 *
## distance  4.73711    0.69632   6.803 1.86e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.038 on 78 degrees of freedom
## Multiple R-squared:  0.8929, Adjusted R-squared:  0.8901
## F-statistic: 325.1 on 2 and 78 DF,  p-value: < 2.2e-16
```

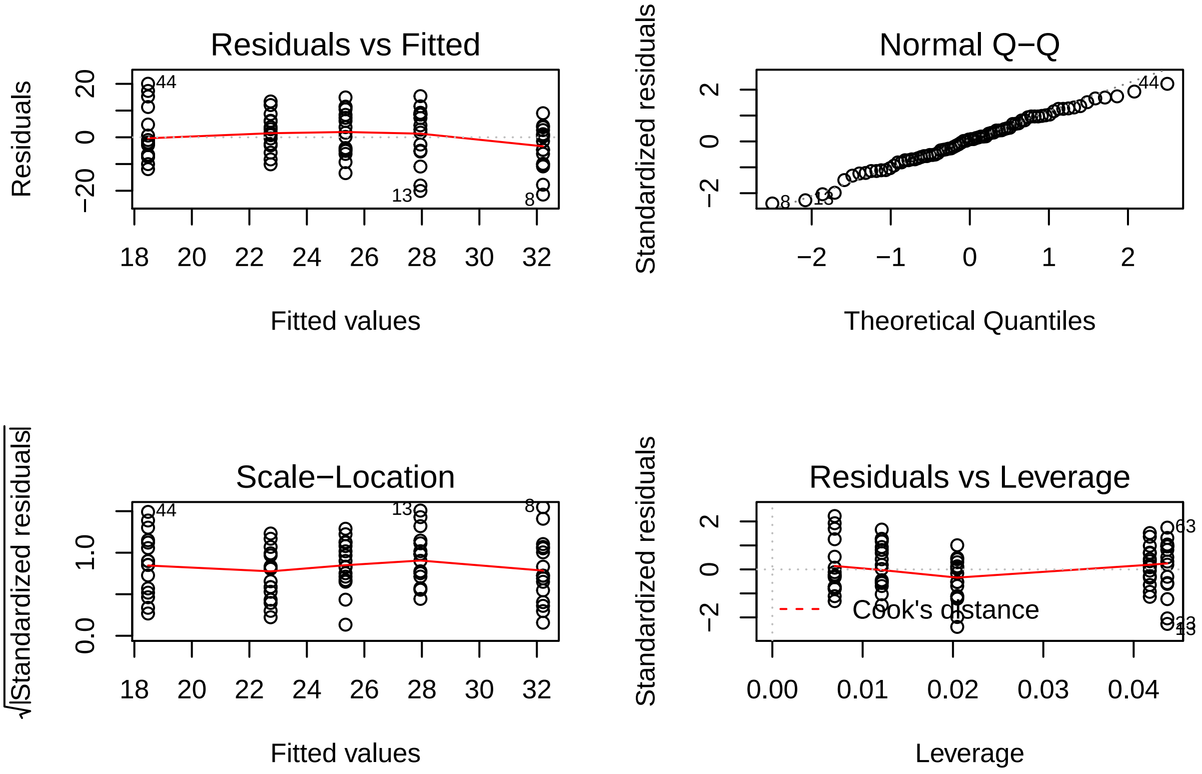
Once again, we performe Lilliefors test and Shapiro-Wilk test to test normality of residuals for the model with transformed response and the Breusch-Pagan test to check heteroskedasticity. Based on the p-values we can not reject the null hypothesis about the normality of the residuals and can not reject the homoskedasticity.

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuals(final_bc.lm_num)
## D = 0.054648, p-value = 0.8038
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(final_bc.lm_num)
## W = 0.99077, p-value = 0.8415
```

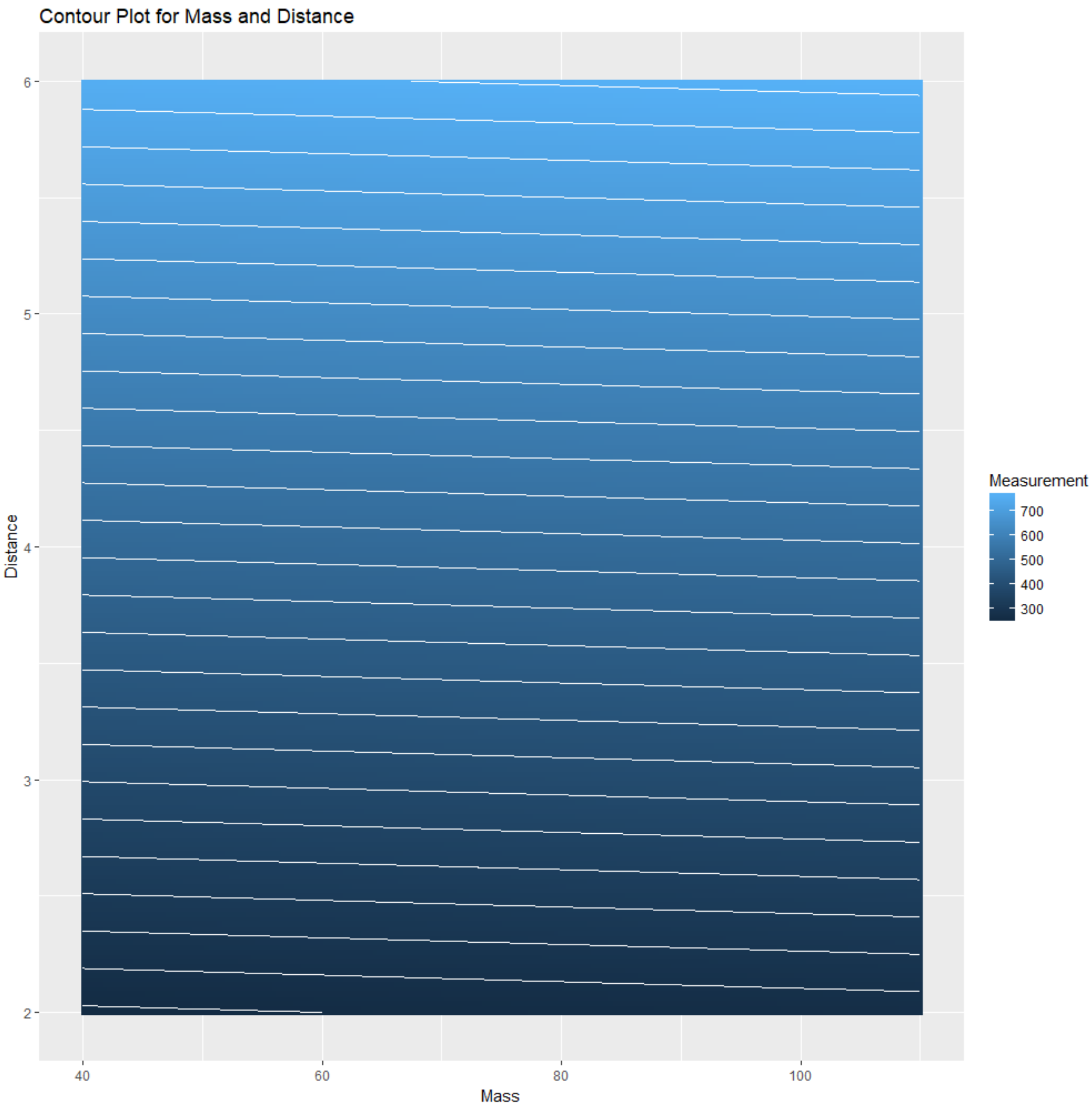
```
##
##  studentized Breusch-Pagan test
##
## data:  final_bc.lm_num
## BP = 2.8518, df = 1, p-value = 0.09127
```

The following Figure provides residual plots for the final model with transformed response varible.



Contour plot

Creating contour plot to show the predictions with repsect to mass and distance values.


Contour Plot for Mass and Distance
Distance
6
5
4
3
2
Mass
40
60
80
100
Measurement
700
600
500
400
300