R Notebook

Дисперсионный анализ. Пример

Загрузим данные (требуется установить Рабочую папку с помощью setwd) или указать полный путь

```
data = read.csv("https://raw.githubusercontent.com/SergeyMirvoda/MD-DA-2017/master/data/diet.csv",row.names=1)
summary(data)
```

```
##      gender            Age            Height        pre.weight
##  Min.   :0.0000   Min.   :16.00   Min.   :141.0   Min.   : 58.00
##  1st Qu.:0.0000   1st Qu.:32.25   1st Qu.:164.2   1st Qu.: 66.00
##  Median :0.0000   Median :39.00   Median :169.5   Median : 72.00
##  Mean   :0.4342   Mean   :39.15   Mean   :170.8   Mean   : 72.53
##  3rd Qu.:1.0000   3rd Qu.:46.75   3rd Qu.:174.8   3rd Qu.: 78.00
##  Max.   :1.0000   Max.   :60.00   Max.   :201.0   Max.   :103.00
##  NA's   :2
##       Diet        weight6weeks
##  Min.   :1.000   Min.   : 53.00
##  1st Qu.:1.000   1st Qu.: 61.85
##  Median :2.000   Median : 68.95
##  Mean   :2.038   Mean   : 68.68
##  3rd Qu.:3.000   3rd Qu.: 73.83
##  Max.   :3.000   Max.   :103.00
##
```

Ознакомимся со структурой и переименуем колонки, как нам удобно
https://www.sheffield.ac.uk/polopoly_fs/1.547015!/file/Diet_data_description.docx (https://www.sheffield.ac.uk/polopoly_fs/1.547015!/file/Diet_data_description.docx)
https://www.sheffield.ac.uk/mash/data (https://www.sheffield.ac.uk/mash/data)

```
colnames(data) <- c("gender", "age", "height", "initial.weight",
                    "diet.type", "final.weight")
data$diet.type <- factor(c("A", "B", "C")[data$diet.type])
summary(data)
```

```
##      gender            age            height      initial.weight
##  Min.   :0.0000   Min.   :16.00   Min.   :141.0   Min.   : 58.00
##  1st Qu.:0.0000   1st Qu.:32.25   1st Qu.:164.2   1st Qu.: 66.00
##  Median :0.0000   Median :39.00   Median :169.5   Median : 72.00
##  Mean   :0.4342   Mean   :39.15   Mean   :170.8   Mean   : 72.53
##  3rd Qu.:1.0000   3rd Qu.:46.75   3rd Qu.:174.8   3rd Qu.: 78.00
##  Max.   :1.0000   Max.   :60.00   Max.   :201.0   Max.   :103.00
##  NA's   :2
##  diet.type  final.weight
##  A:24       Min.   : 53.00
##  B:27       1st Qu.: 61.85
##  C:27       Median : 68.95
##             Mean   : 68.68
##             3rd Qu.: 73.83
##             Max.   :103.00
##
```

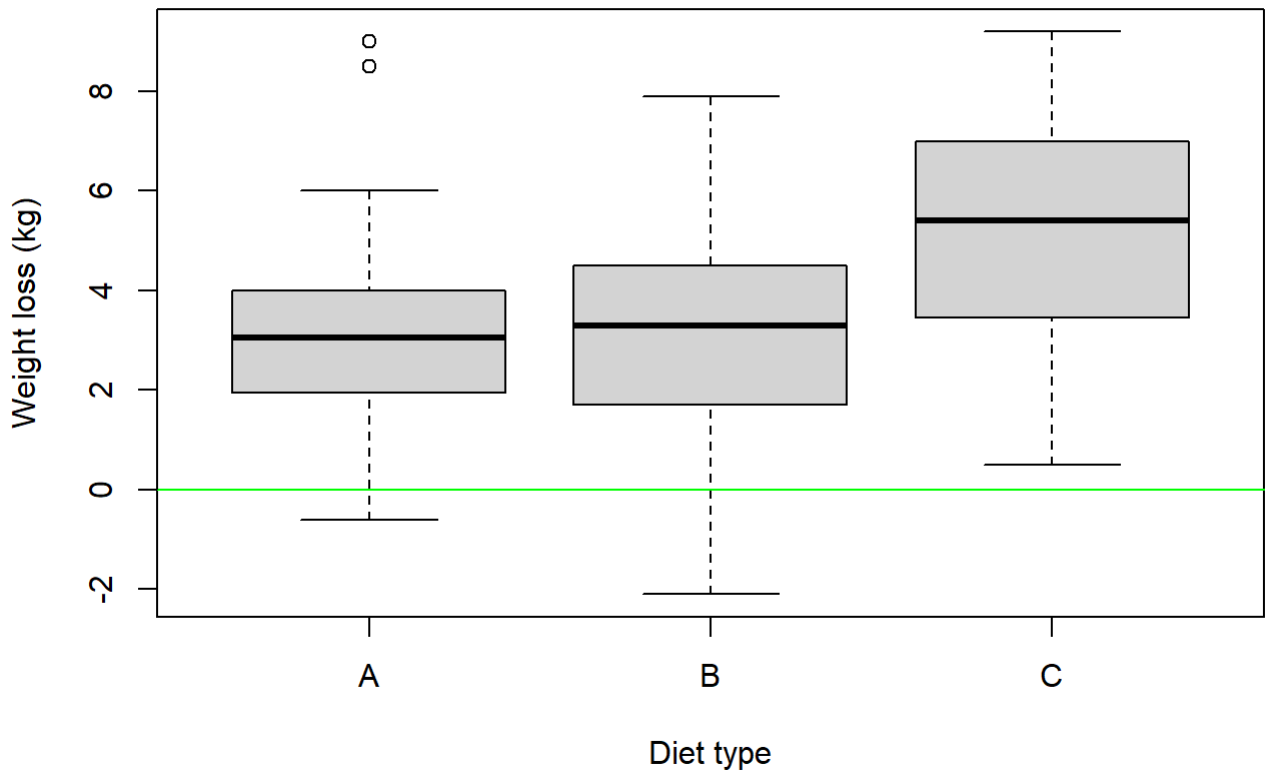
Добавим новую колонку - Похудение

```
data$weight.loss = data$initial.weight - data$final.weight
summary(data)
```

```
##      gender            age            height      initial.weight
##  Min.   :0.0000   Min.   :16.00   Min.   :141.0   Min.   : 58.00
##  1st Qu.:0.0000   1st Qu.:32.25   1st Qu.:164.2   1st Qu.: 66.00
##  Median :0.0000   Median :39.00   Median :169.5   Median : 72.00
##  Mean   :0.4342   Mean   :39.15   Mean   :170.8   Mean   : 72.53
##  3rd Qu.:1.0000   3rd Qu.:46.75   3rd Qu.:174.8   3rd Qu.: 78.00
##  Max.   :1.0000   Max.   :60.00   Max.   :201.0   Max.   :103.00
##  NA's   :2
##  diet.type  final.weight     weight.loss
##  A:24       Min.   : 53.00   Min.   :-2.100
##  B:27       1st Qu.: 61.85   1st Qu.: 2.000
##  C:27       Median : 68.95   Median : 3.600
##             Mean   : 68.68   Mean   : 3.845
##             3rd Qu.: 73.83   3rd Qu.: 5.550
##             Max.   :103.00   Max.   : 9.200
##
```

Проанализируем есть ли различия по типам диет

```
boxplot(weight.loss~diet.type,data=data,col="light gray",
        ylab = "Weight loss (kg)", xlab = "Diet type")
abline(h=0,col="green")
```



проверим сбалансированные ли данные

```
table(data$diet.type)
```

```
##
##  A  B  C
## 24 27 27
```

График групповых средних

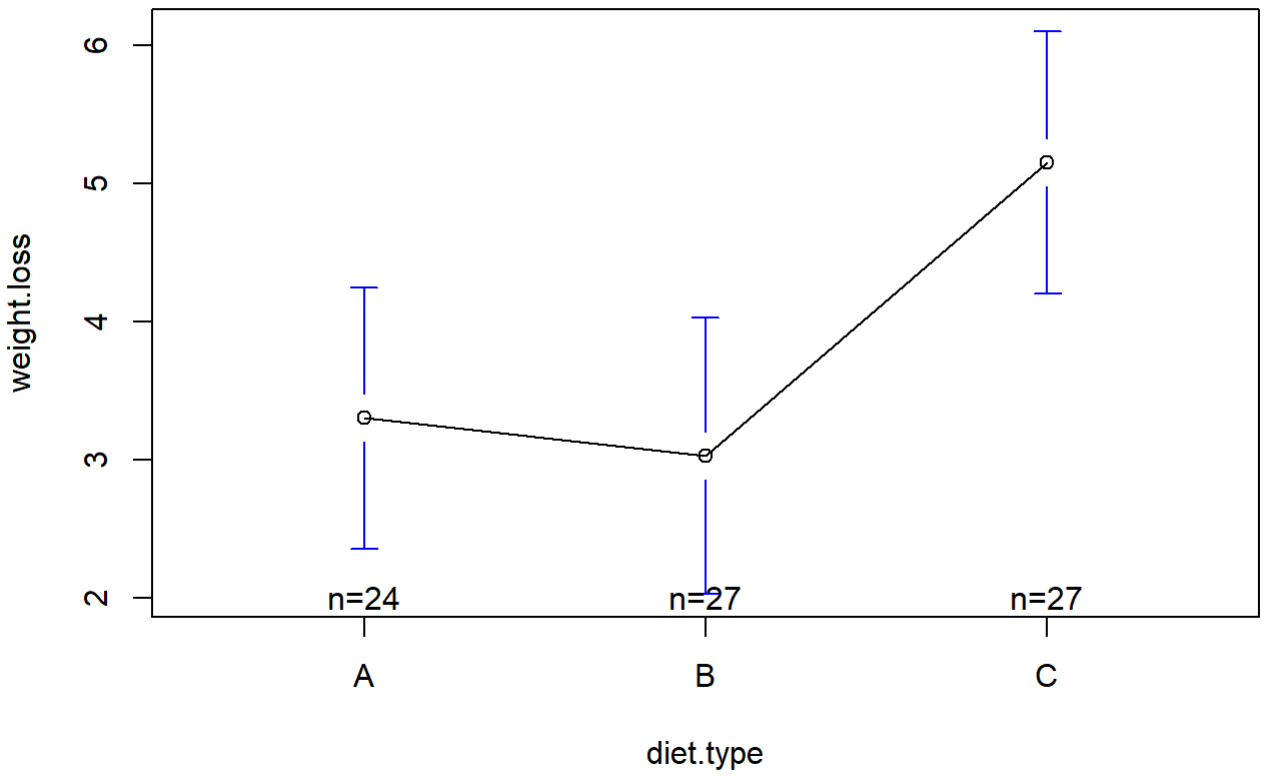
```
library(gplots) #библиотека устанавлевается с помощью install.packages
```

```
## Warning: package 'gplots' was built under R version 3.4.3
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
plotmeans(weight.loss ~ diet.type, data=data)
```




```
aggregate(data$weight.loss, by = list(data$diet.type), FUN=sd)
```

Group.1 <fctr>	x <dbl>
A	2.240148
B	2.523367
C	2.395568

3 rows

Для подгонки ANOVA модели используем функцию aov, частный случай линейной модели lm тест на межгрупповые различия

```
fit <- aov(weight.loss ~ diet.type, data=data)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## diet.type    2   71.1   35.55   6.197 0.00323 **
## Residuals   75  430.2    5.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

попарные различия между средними значениями для всех групп

```
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight.loss ~ diet.type, data = data)
##
## $diet.type
##           diff        lwr      upr     p adj
## B-A -0.2740741 -1.8806155 1.332467 0.9124737
## C-A  1.8481481  0.2416067 3.454690 0.0201413
## C-B  2.1222222  0.5636481 3.680796 0.0047819
```

Tukey honest significant differences test)

```
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 3.4.3
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

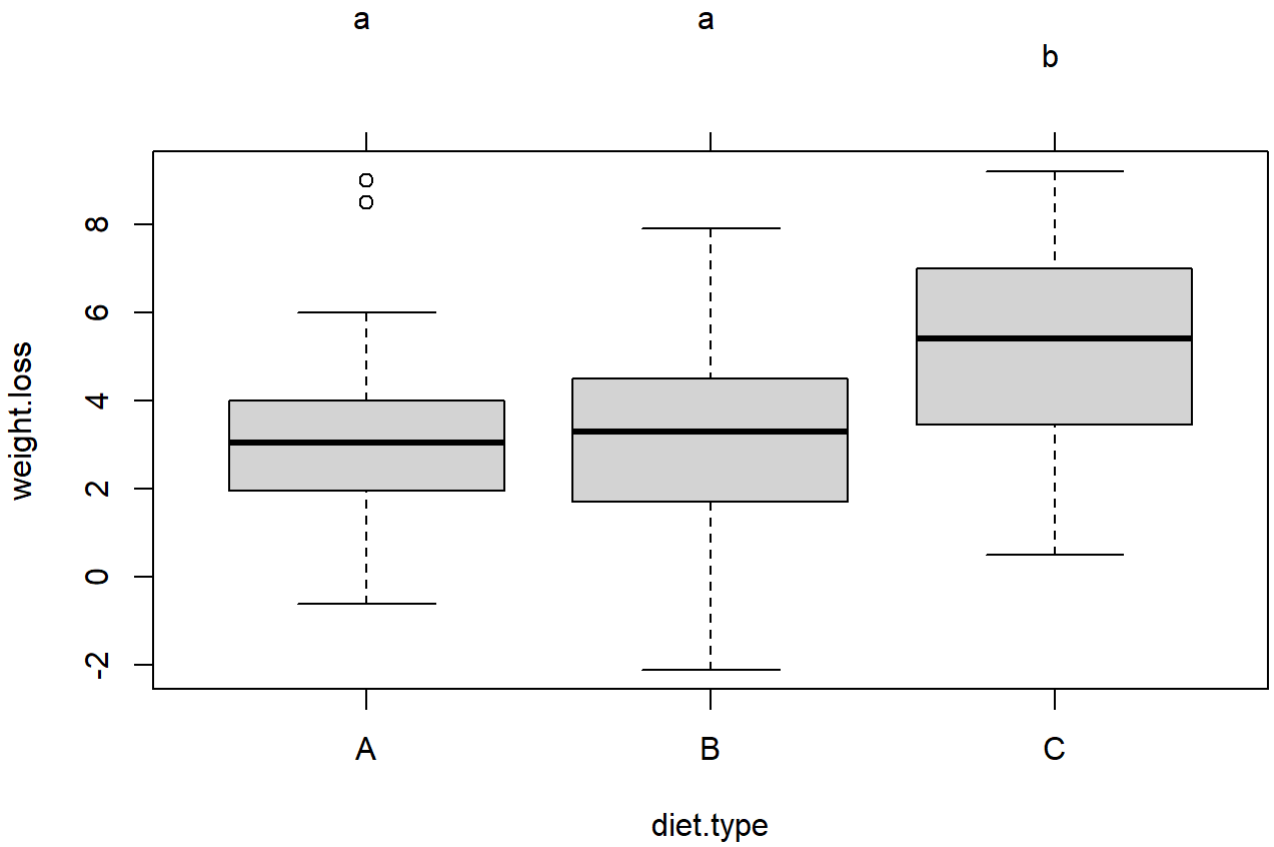
```
## Warning: package 'TH.data' was built under R version 3.4.3
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'
```

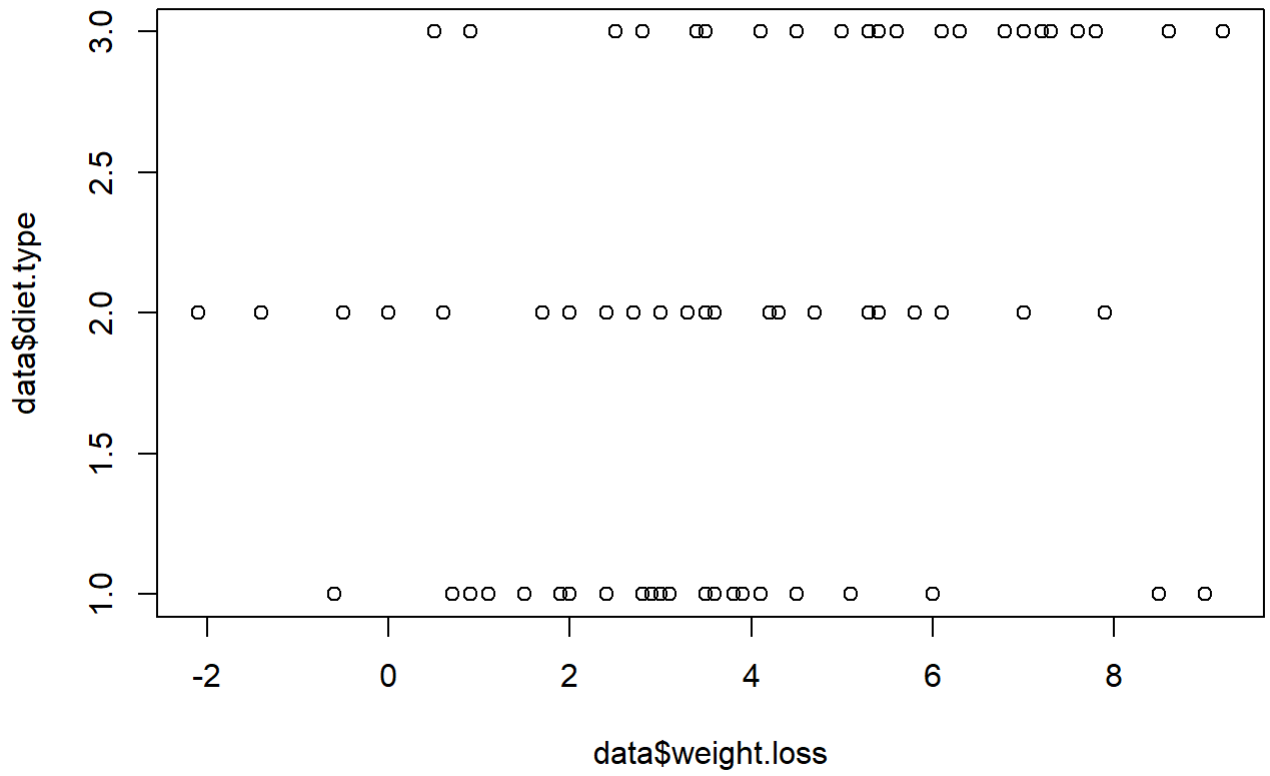
```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
par(mar=c(5,4,6,2))
tuk <- glht(fit, linfct=mcp(diet.type="Tukey"))
plot(cld(tuk, level=.05),col="lightgrey")
```



###Задание ###Добавить проверку на выборы и избавиться от них

```
plot(data$weight.loss,data$diet.type)
```




```
data.noout<-data[data$weight.loss<=8&data$weight.loss>=0,]
```

повторно проверсти все тесты и сравнить результаты с выбросами и без

Различия по типам диет

```
boxplot(weight.loss~diet.type,data=data.noout,col="light gray",
        ylab = "Weight loss (kg)", xlab = "Diet type")
abline(h=0,col="green")
```

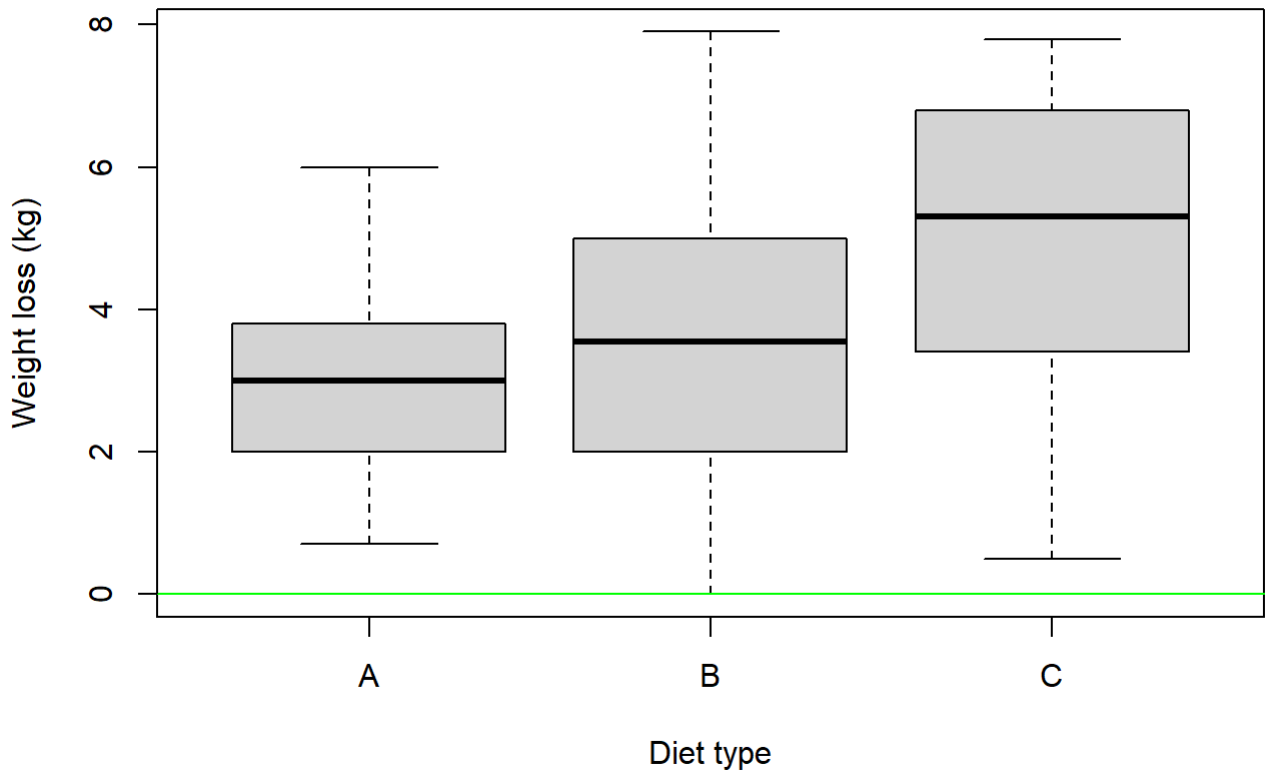
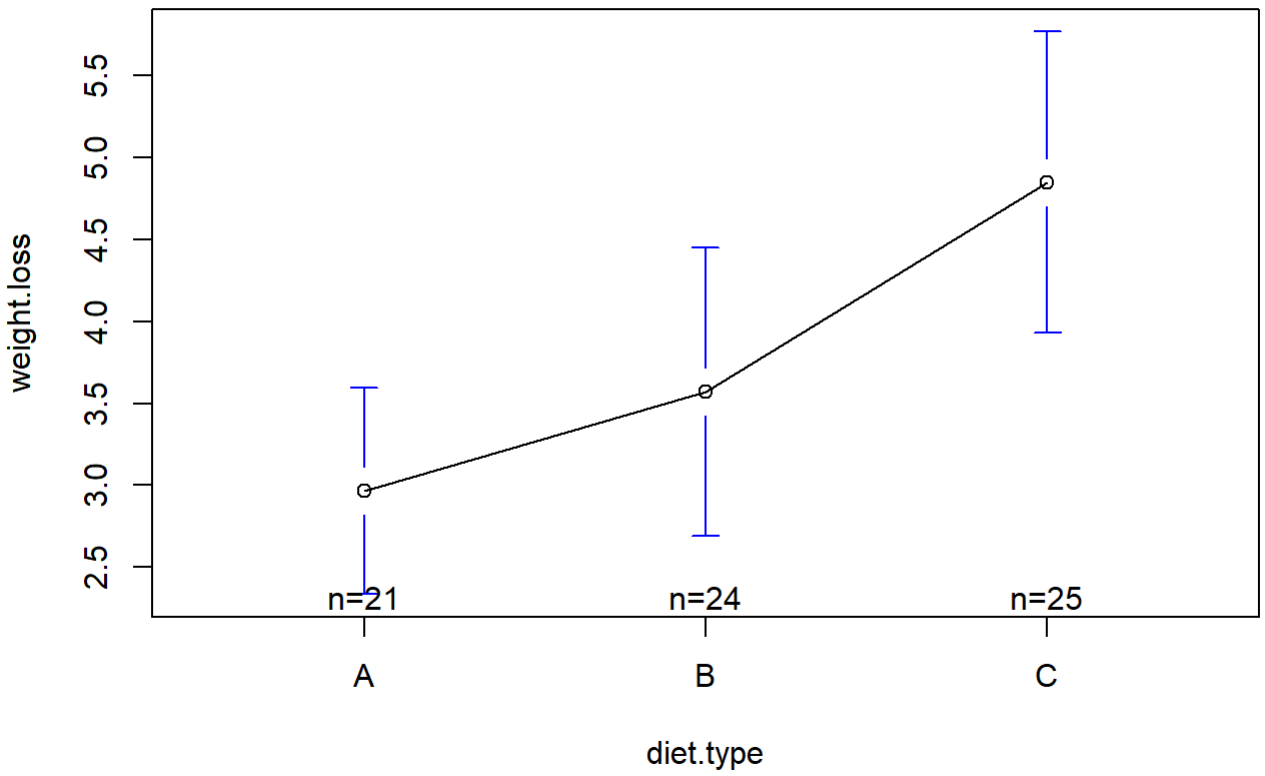


График групповых средних

```
library(gplots)
plotmeans(weight.loss ~ diet.type, data=data.noout)
```




```
aggregate(data$weight.loss, by = list(data$diet.type), FUN=sd)
```

Group.1 <fctr>	x <dbl>
A	2.240148
B	2.523367
C	2.395568

3 rows

Для подгонки ANOVA модели используем функцию aov, частный случай линейной модели lm

```
#тест на межгрупповые различия
fit.noout <- aov(weight.loss ~ diet.type, data=data.noout)
summary(fit.noout)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## diet.type    2  43.16  21.578   5.623 0.00553 **
## Residuals   67 257.10   3.837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

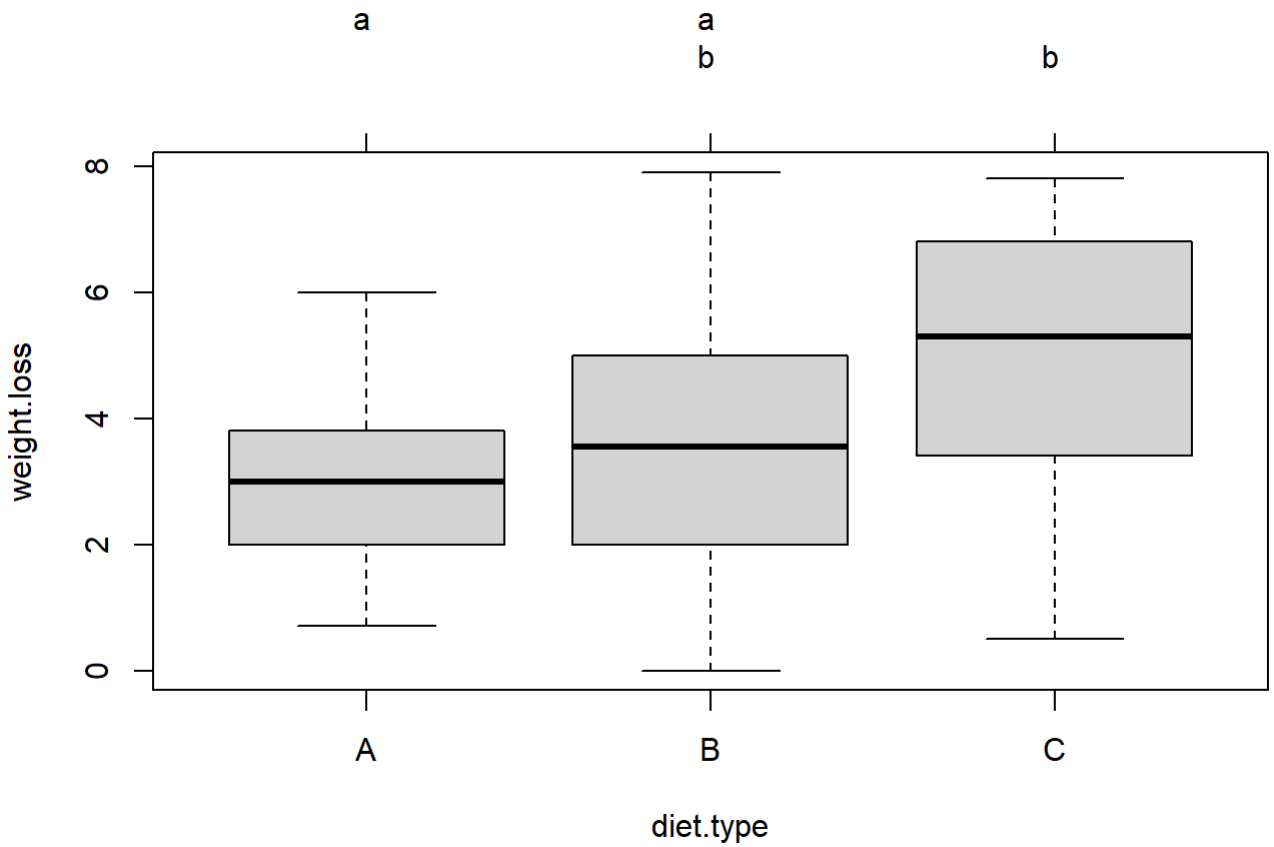
Попарные различия между средними значениями для всех групп

```
TukeyHSD(fit.noout)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight.loss ~ diet.type, data = data.noout)
##
## $diet.type
##          diff         lwr      upr     p adj
## B-A 0.6041667 -0.79880872 2.007142 0.5593442
## C-A 1.8813333  0.49151406 3.271153 0.0051569
## C-B 1.2771667 -0.06461384 2.618947 0.0653789
```

Tukey honest significant differences test

```
library(multcomp)
par(mar=c(5,4,6,2))
tuk <- glht(fit.noout, linfct=mcp(diet.type="Tukey"))
plot(cld(tuk, level=.05),col="lightgrey")
```



Диета С лучше А и В ###Открыть документ
https://www.sheffield.ac.uk/polopoly_fs/1.547015!/file/Diet_data_description.docx

(https://www.sheffield.ac.uk/polopoly_fs/1.547015!/file/Diet_data_description.docx) ###и попытаться выполнить задания из него Зависимость потери веса от пола

```
data.noout.with.gender<-data[!is.na(data$gender),]
data.noout.with.gender$gender <- factor(c("Female","Male")[as.ordered(data.noout.with.gender$gender)])
boxplot(weight.loss~gender,data=data.noout.with.gender,col="light gray",
        ylab = "Weight loss (kg)", xlab = "Gender")
abline(h=0,col="green")
```

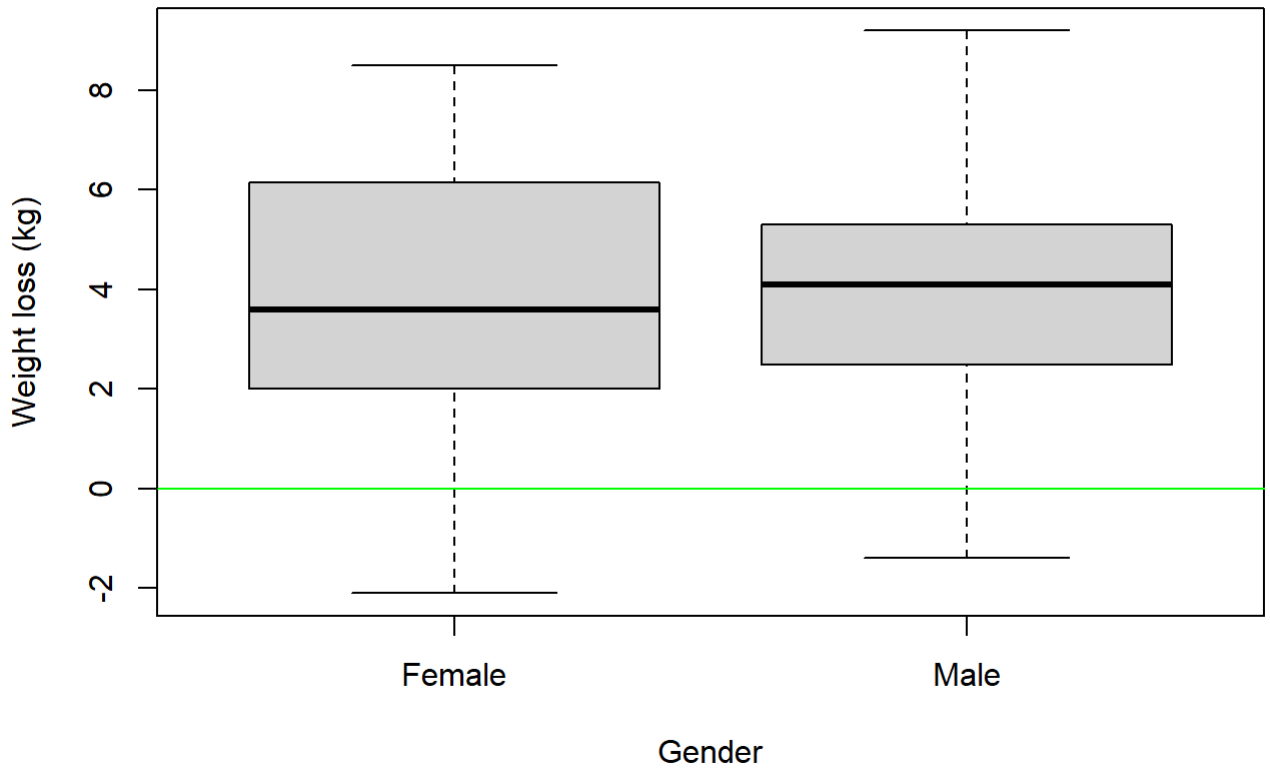
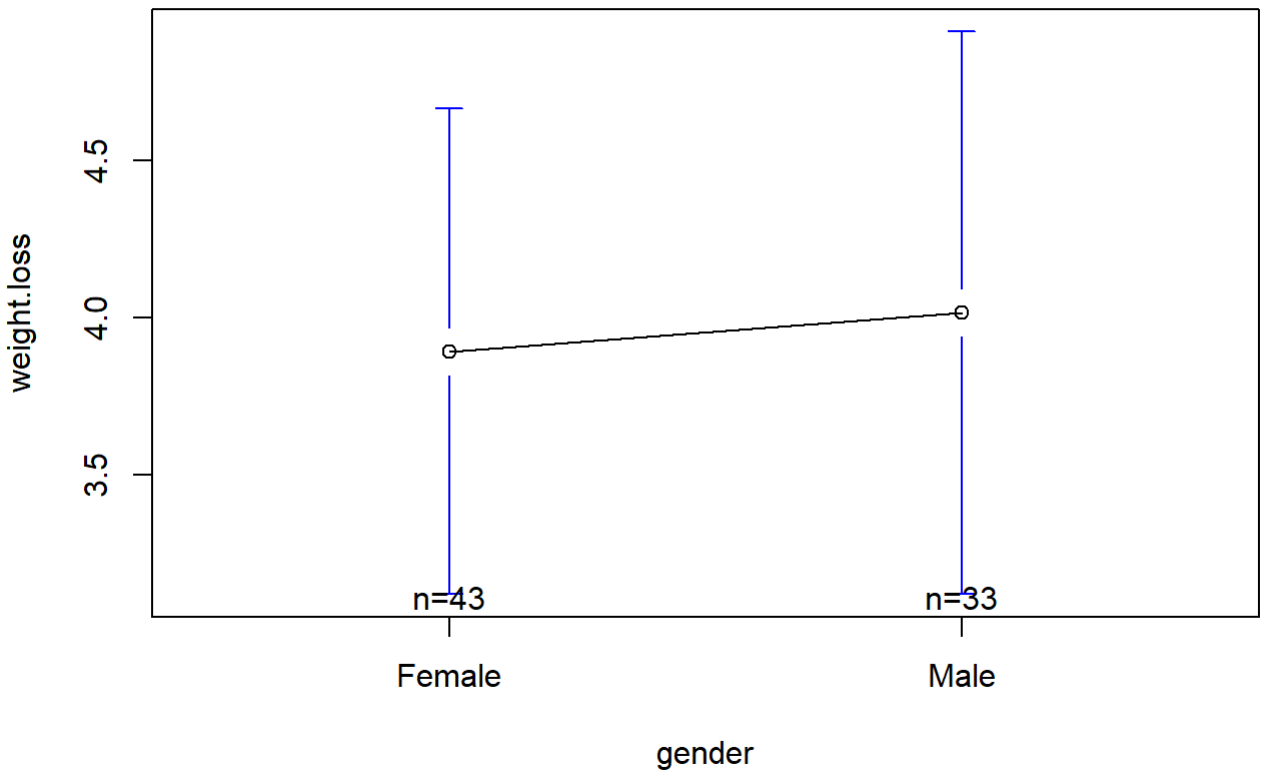


График групповых средних

```
plotmeans(weight.loss ~ gender, data=data.noout.with.gender)
```




```
aggregate(data.noout.with.gender$weight.loss, by = list(data.noout.with.gender$gender), FUN=sd)
```

Group.1 <fctr>	**x** <dbl>
Female	2.515892
Male	2.529837

2 rows

Для подгонки ANOVA модели используем функцию aov, частный случай линейной модели lm тест на межгрупповые различия

```
fit.noout <- aov(weight.loss ~ gender, data=data.noout.with.gender)
summary(fit.noout)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## gender       1    0.3   0.278   0.044  0.835
## Residuals   74  470.7   6.360
```

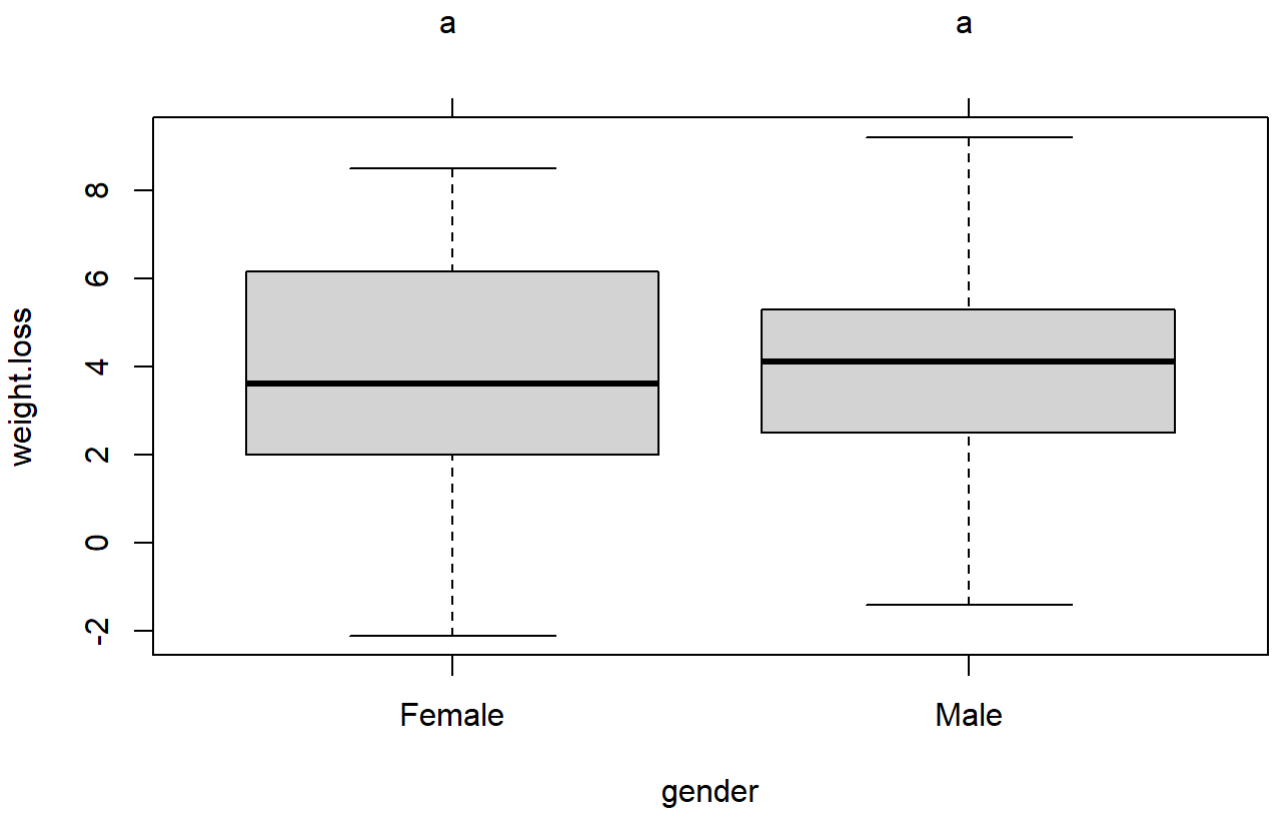
Попарные различия между средними значениями для всех групп

```
TukeyHSD(fit.noout)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight.loss ~ gender, data = data.noout.with.gender)
##
## $gender
##                   diff      lwr      upr     p adj
## Male-Female 0.1221283 -1.04081 1.285067 0.8348274
```

Tukey honest significant differences test

```
library(multcomp)
par(mar=c(5,4,6,2))
tuk <- glht(fit.noout, linfct=mcp(gender="Tukey"))
plot(cld(tuk, level=.05),col="lightgrey")
```



Нельзя сказать, что потеря веса зависит от пола