# Дисперсійний аналіз

Daria Kravets

Завантажуємо датасет:

```
employee1 = read.csv("E://DownloadsE//WA_Fn-UseC_-HR-Employee-Attrition (1).csv")
```

***MonthlyIncome*** - щомісячний дохід; ***JobRole*** - посада.

Підключаємо бібліотеку ***ggplot2***

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

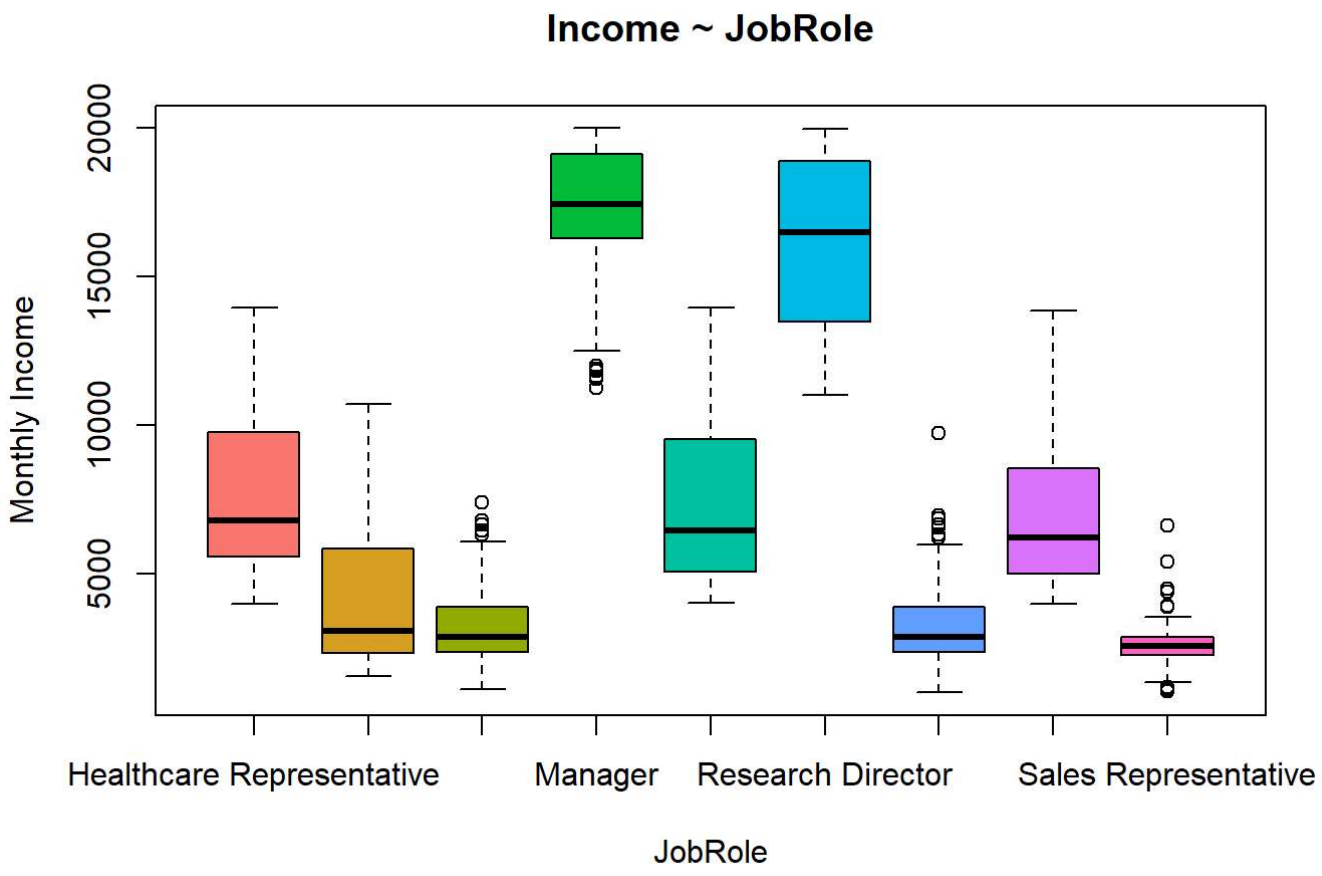
Задаємо змінну ***JobRole*** як фактор:

```
employee1$JobRole <- factor(employee1$JobRole)
```

Подивимось на зображення даних:

```
p<-ggplot(employee1, aes(x=MonthlyIncome, y=JobRole, color=JobRole)) +
  geom_jitter(position=position_jitter(0.2))
p + theme(legend.position="none")
```

```
boxplot(MonthlyIncome ~ JobRole, data = employee1,
        xlab = "JobRole", ylab = "Monthly Income",
        main = "Income ~ JobRole", col = c("#f8766d", "#d69f22", "#93aa00", "#00ba38", "#00c1
9f", "#00b9e3", "#619cff", "#db72fb", "#ff61c3"))
```




Проведемо дисперсійний аналіз:

```
aggregate(x = employee1$MonthlyIncome, by = list(employee1$JobRole), FUN = mean)
```

```
##                    Group.1         x
## 1 Healthcare Representative  7528.763
## 2           Human Resources  4235.750
## 3     Laboratory Technician  3237.170
## 4                   Manager 17181.676
## 5    Manufacturing Director  7295.138
## 6         Research Director 16033.550
## 7        Research Scientist  3239.973
## 8           Sales Executive  6924.279
## 9      Sales Representative  2626.000
```

```
mod1 <- aov(MonthlyIncome ~ JobRole, data = employee1)
summary(aov(MonthlyIncome ~ JobRole, data = employee1))
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## JobRole        8 2.657e+10 3.321e+09   810.2 <2e-16 ***
## Residuals   1461 5.989e+09 4.099e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Гіпотезу про відсутність впливу посади на дохід відхиляємо.

```
summary(lm(MonthlyIncome ~ JobRole, data = employee1))
```

```
##
## Call:
## lm(formula = MonthlyIncome ~ JobRole, data = employee1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5938  -1209   -351   1165   6948
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     7528.8      176.9  42.560  < 2e-16 ***
## JobRoleHuman Resources         -3293.0      331.9  -9.923  < 2e-16 ***
## JobRoleLaboratory Technician   -4291.6      217.1 -19.770  < 2e-16 ***
## JobRoleManager                  9652.9      267.4  36.104  < 2e-16 ***
## JobRoleManufacturing Director   -233.6      244.1  -0.957  0.33860
## JobRoleResearch Director        8504.8      287.3  29.604  < 2e-16 ***
## JobRoleResearch Scientist      -4288.8      212.9 -20.143  < 2e-16 ***
## JobRoleSales Executive          -604.5      209.4  -2.886  0.00396 **
## JobRoleSales Representative    -4902.8      284.0 -17.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2025 on 1461 degrees of freedom
## Multiple R-squared:  0.8161, Adjusted R-squared:  0.8151
## F-statistic: 810.2 on 8 and 1461 DF,  p-value: < 2.2e-16
```

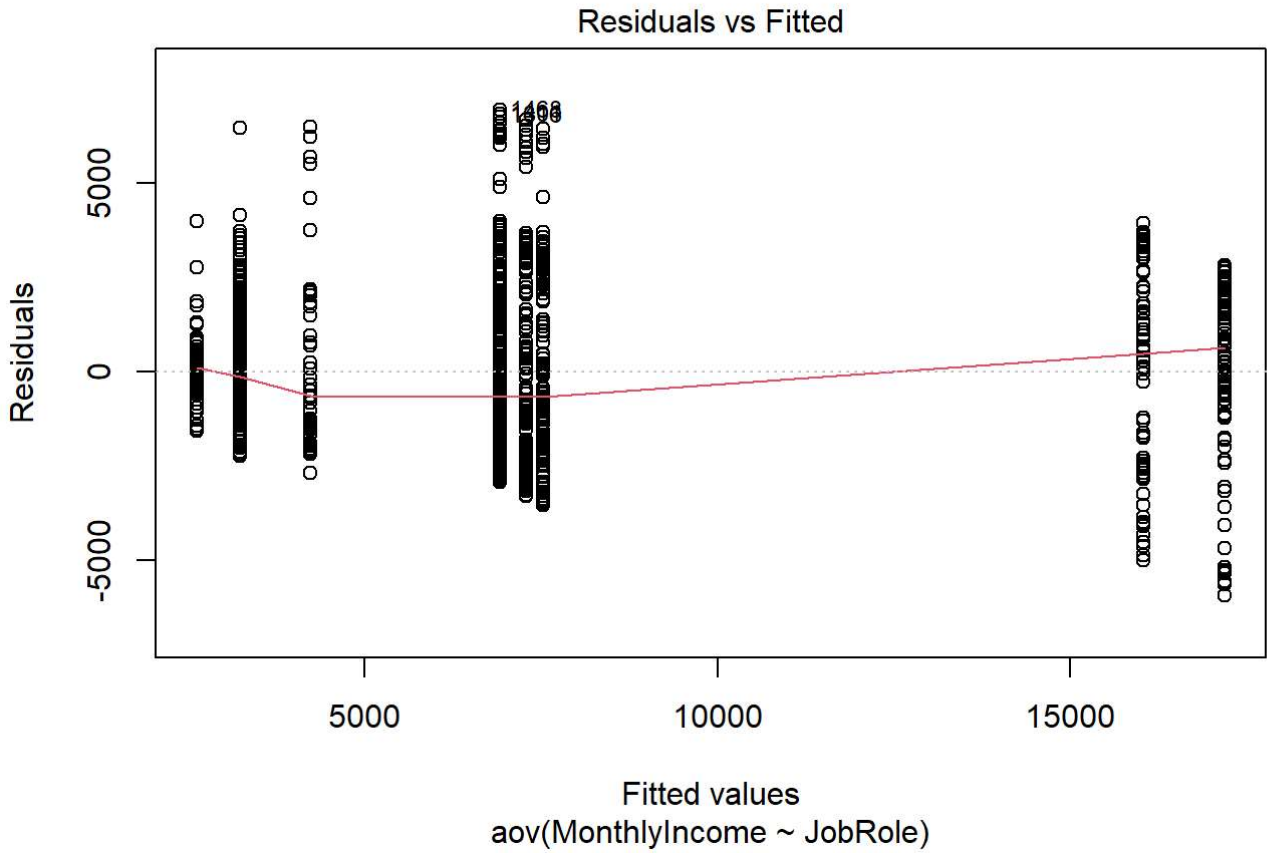
```
contrasts(employee1$JobRole) <- contr.sum
contrasts(employee1$JobRole)
```

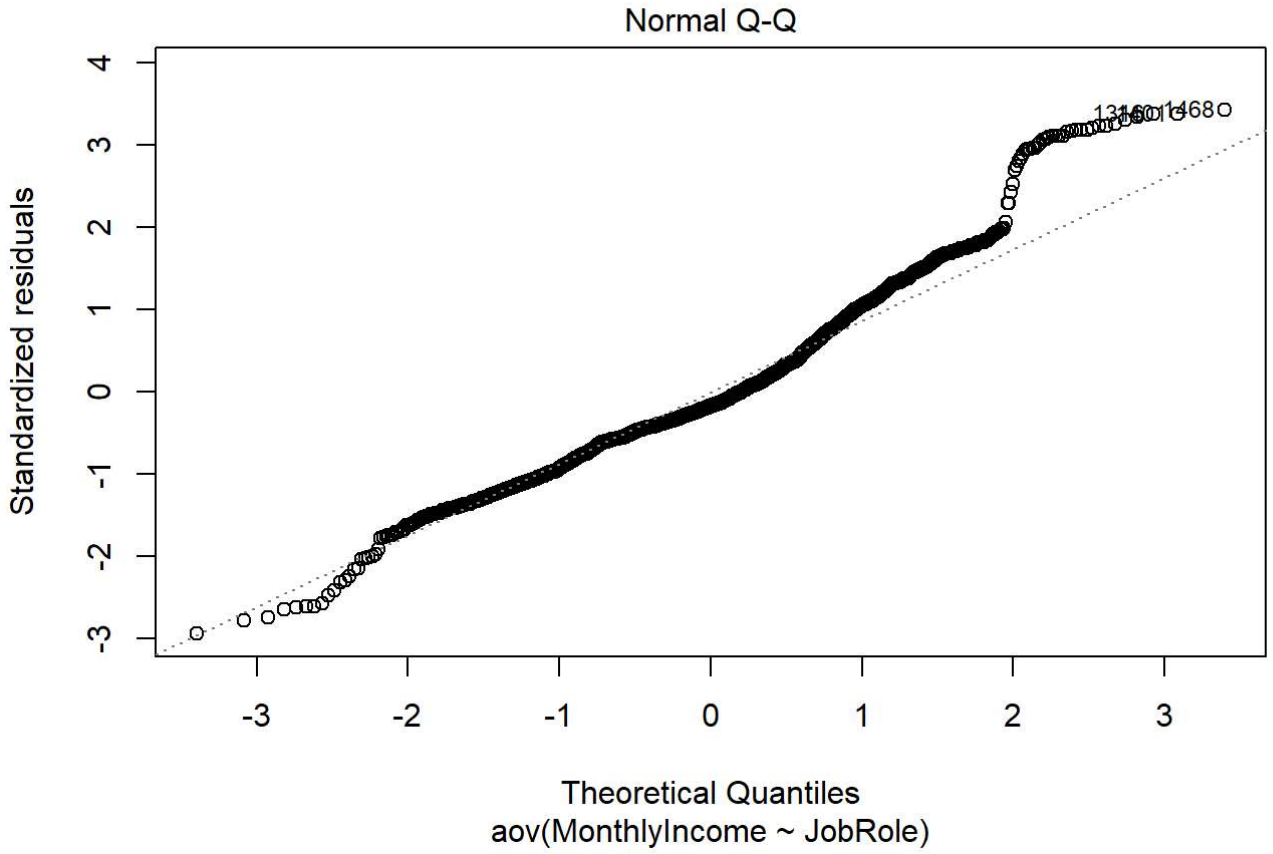
```
##                           [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## Healthcare Representative    1    0    0    0    0    0    0    0
## Human Resources              0    1    0    0    0    0    0    0
## Laboratory Technician        0    0    1    0    0    0    0    0
## Manager                      0    0    0    1    0    0    0    0
## Manufacturing Director       0    0    0    0    1    0    0    0
## Research Director            0    0    0    0    0    1    0    0
## Research Scientist           0    0    0    0    0    0    1    0
## Sales Executive              0    0    0    0    0    0    0    1
## Sales Representative        -1   -1   -1   -1   -1   -1   -1   -1
```

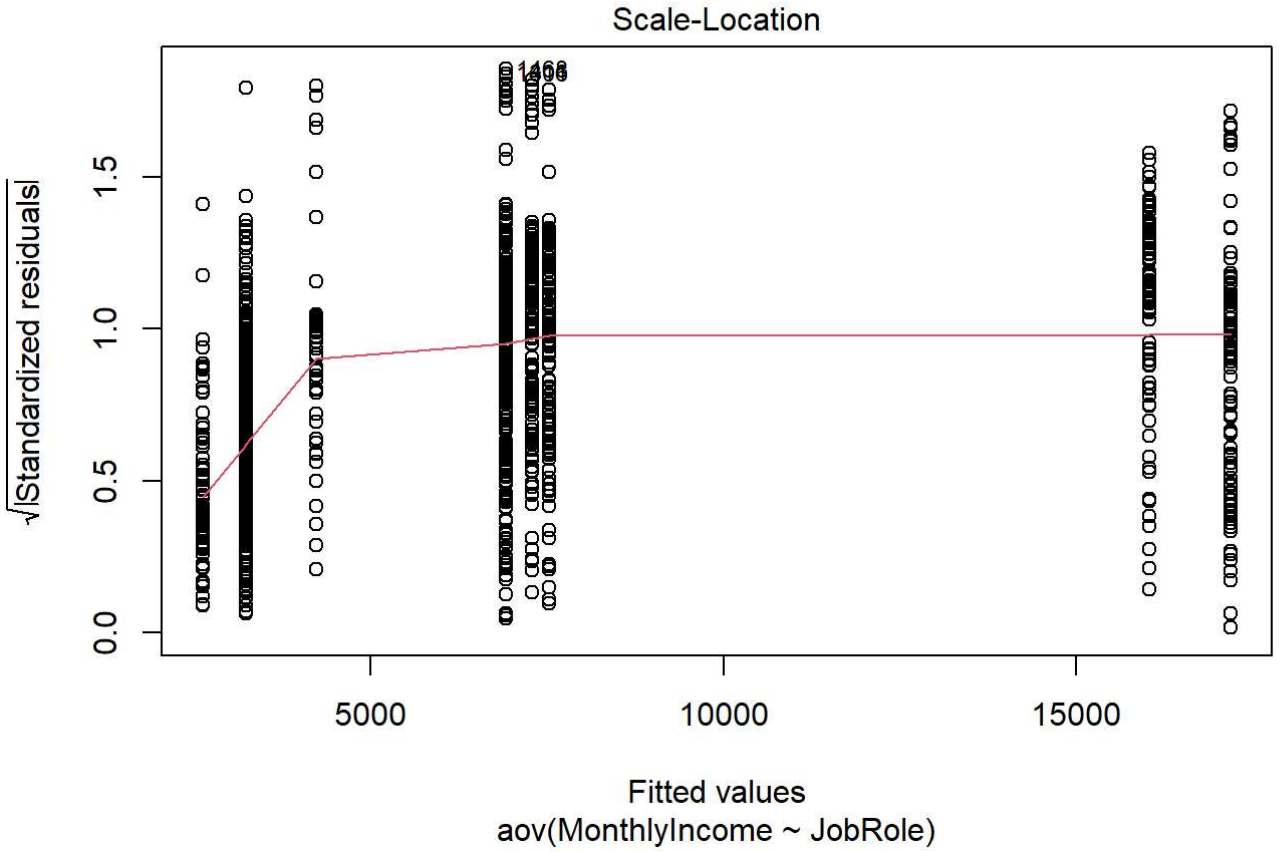
```
summary(lm(MonthlyIncome ~ JobRole, data = employee1))
```

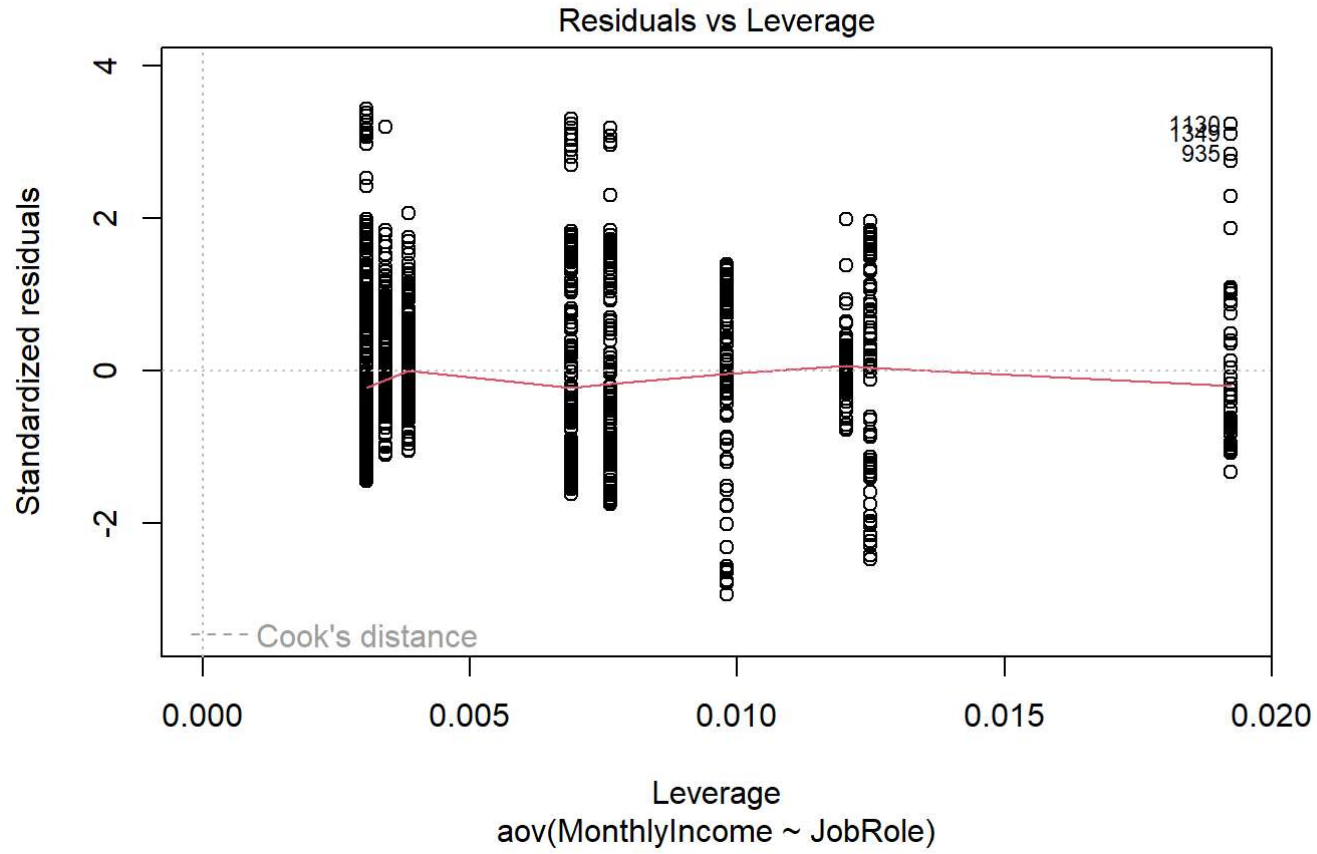
```
##
## Call:
## lm(formula = MonthlyIncome ~ JobRole, data = employee1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5938  -1209   -351   1165   6948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7589.14      63.02 120.430  < 2e-16 ***
## JobRole1      -60.38     168.26  -0.359   0.7197
## JobRole2    -3353.39     255.51 -13.124  < 2e-16 ***
## JobRole3    -4351.97     127.60 -34.107  < 2e-16 ***
## JobRole4     9592.53     187.70  51.107  < 2e-16 ***
## JobRole5     -294.01     161.12  -1.825   0.0682 .
## JobRole6     8444.41     209.35  40.337  < 2e-16 ***
## JobRole7    -4349.17     122.03 -35.641  < 2e-16 ***
## JobRole8     -664.87     117.27  -5.670 1.72e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2025 on 1461 degrees of freedom
## Multiple R-squared:  0.8161, Adjusted R-squared:  0.8151
## F-statistic: 810.2 on 8 and 1461 DF,  p-value: < 2.2e-16
```

```
plot(mod1)
```

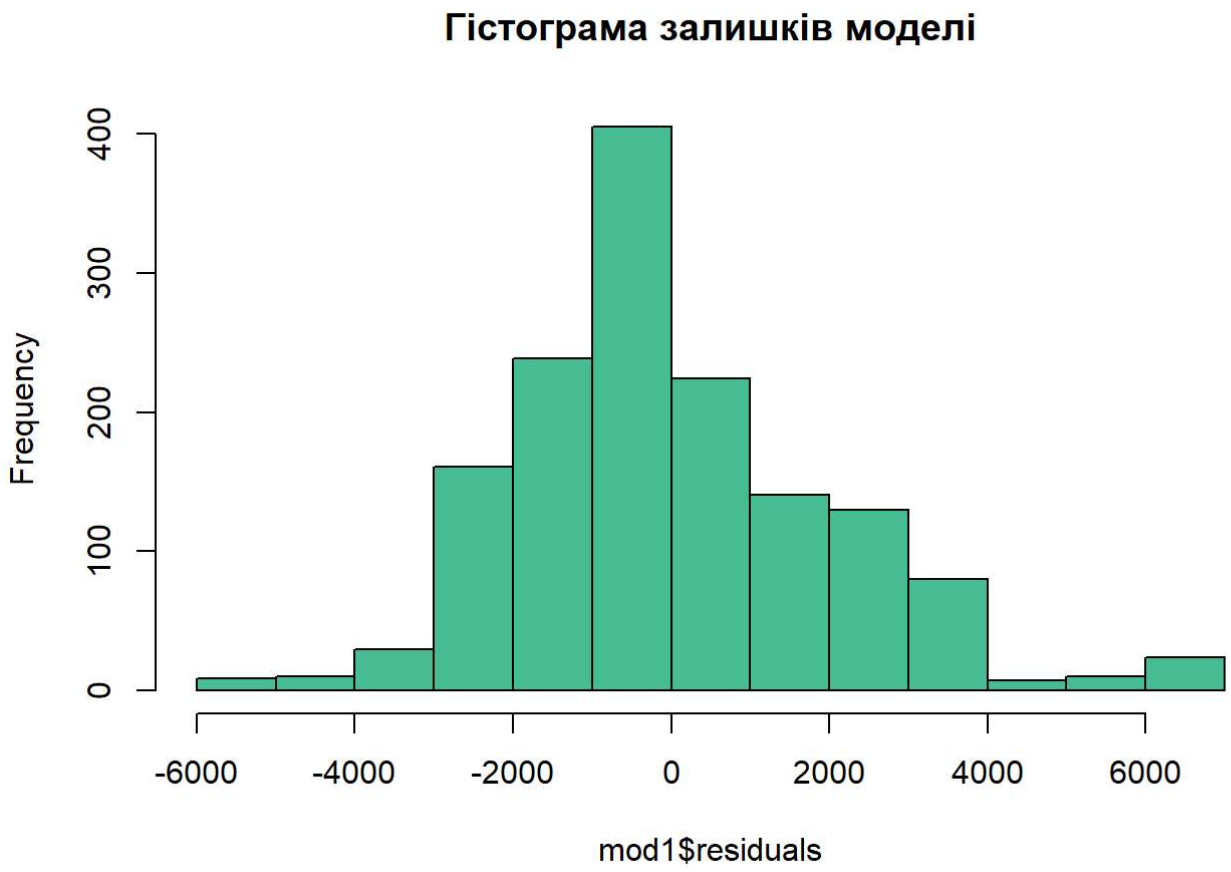

Residuals vs Fitted
Residuals
5000
0
-5000
1468
5000
10000
15000
Fitted values
aov(MonthlyIncome ~ JobRole)


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
-3
1468
-3
-2
-1
0
1
2
3
Theoretical Quantiles
aov(MonthlyIncome ~ JobRole)




```
hist(mod1$residuals, col = "#48bd94", main="Гістограма залишків моделі")
```



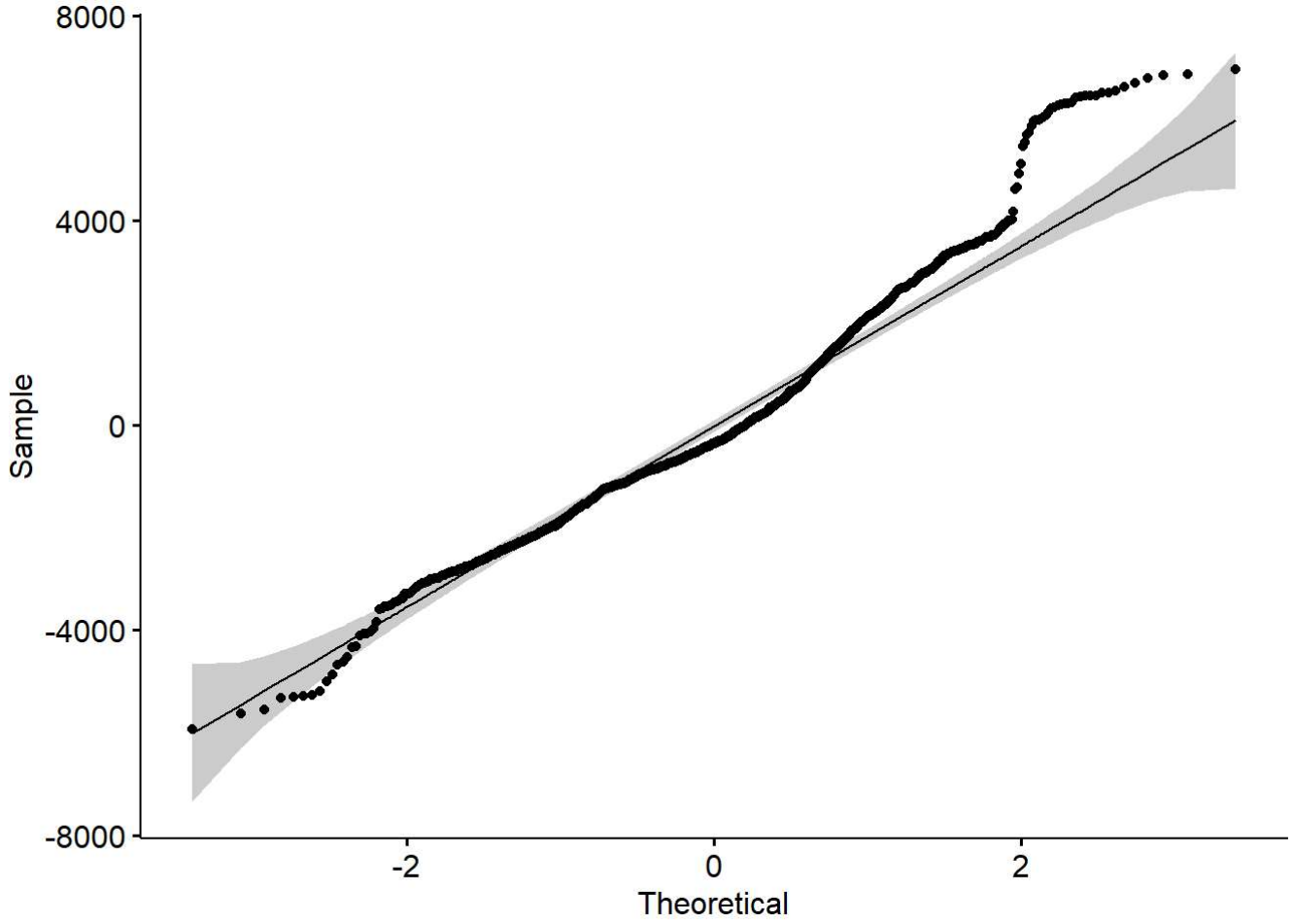

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.2.2
```

```
ggqqplot(mod1$residuals)
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
shapiro.test(mod1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod1$residuals
## W = 0.96705, p-value < 2.2e-16
```

```
tapply(employee1$MonthlyIncome, employee1$JobRole, var)
```

```
## Healthcare Representative           Human Resources     Laboratory Technician
##                 6464561.4                 5947988.1                1323074.6
##                   Manager    Manufacturing Director         Research Director
##                 5369415.0                 7164967.8                7995442.6
##        Research Scientist           Sales Executive      Sales Representative
##                 1435482.2                 5602419.3                 730229.1
```

```
bartlett.test(employee1$MonthlyIncome, employee1$JobRole)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  employee1$MonthlyIncome and employee1$JobRole
## Bartlett's K-squared = 397.13, df = 8, p-value < 2.2e-16
```