anova_no_zero

Gulzina Kuttubekova

12/5/2019

```
games <- read.csv('games.csv')
sales <- read.csv('sales.csv')
```

```
# log-transform and remove zeros
sales_nonzero <- sales %>% select_all() %>%
    filter(Sales != 0)
```

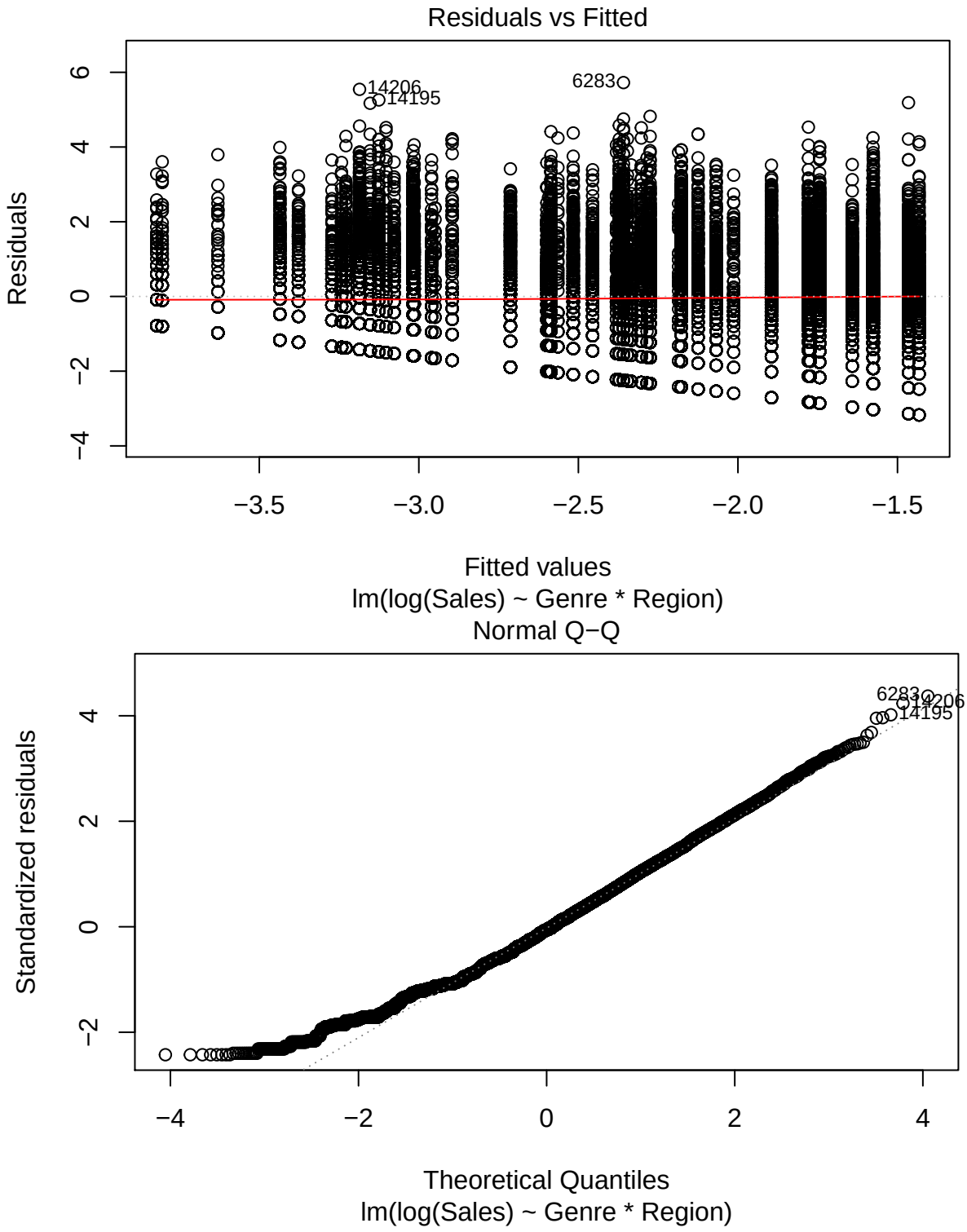
Hypothesis 1:

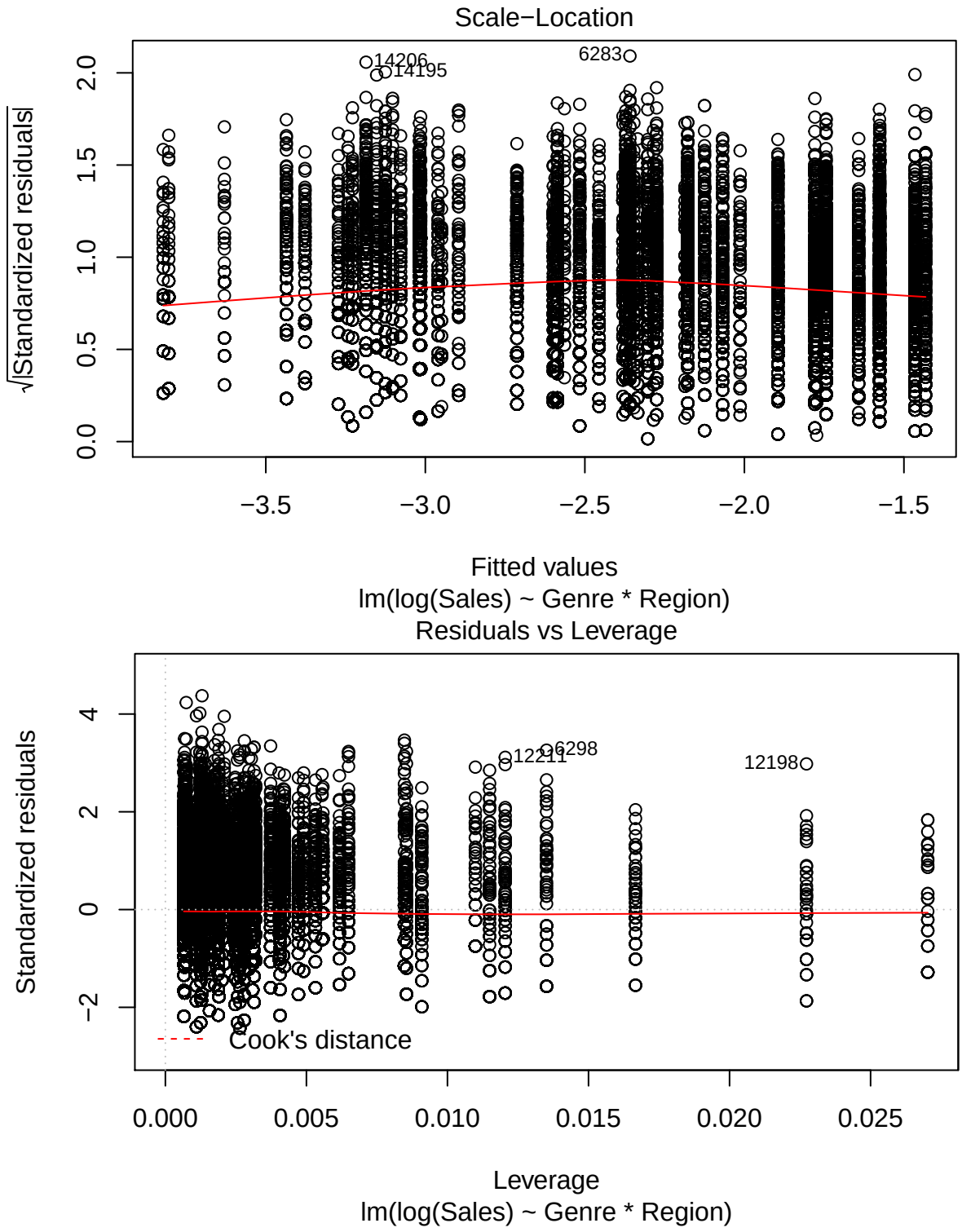
```
genre_region <- lm(log(Sales) ~ Genre*Region, data = sales_nonzero)
genre_region %>% anova()
```

```
## Analysis of Variance Table
##
## Response: log(Sales)
##                 Df Sum Sq Mean Sq   F value    Pr(>F)
## Genre           11    661   60.08   35.0251 < 2.2e-16 ***
## Region           3   6857 2285.55 1332.5056 < 2.2e-16 ***
## Genre:Region    33    511   15.48    9.0269 < 2.2e-16 ***
## Residuals    19746  33869    1.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check the assumptions:

```
plot(genre_region)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(Sales) ~ Genre * Region)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(Sales) ~ Genre * Region)
14206
14195
6283



Based on the QQ plot and residual plot there isn't much deviation from the normality and homogenity assumptions. There appears to be a small pattern in residual plot, we will check if it's significant by additional tests:

```
shapiro.test(genre_region$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  genre_region$residuals[sample(1:length(sales_nonzero$Platform),     5000)]
## W = 0.98956, p-value < 2.2e-16
```

We see that the normality assumption is actually violated. What about homogeneity? - It's also violated!!! There are no outliers and influential points.

```
bptest(genre_region)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  genre_region
## BP = 426.17, df = 47, p-value < 2.2e-16
```

WLS method might be a potential remedy for the assumptions violations. So try this:

```
varfunc_gen_reg <- lm(log(genre_region$residuals^2) ~ Genre*Region,
                      data = sales_nonzero)
sales_nonzero$varfunc_gen_reg <- exp(varfunc_gen_reg$fitted.values)
genre_region_gls <- lm(log(Sales) ~ Genre*Region,
                       weights = 1/sqrt(varfunc_gen_reg),
                       data = sales_nonzero)
```

Check normality:

```
shapiro.test(genre_region_gls$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  genre_region_gls$residuals[sample(1:length(sales_nonzero$Platform),     5000)]
## W = 0.98945, p-value < 2.2e-16
```

P-value had increased from previous LS case, but still is less than 0.05. Hence WLS method didn't help to resolve the abnormality :D issue.

```
bptest(genre_region_gls)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  genre_region_gls
## BP = 426.17, df = 47, p-value < 2.2e-16
```

Homogeneity is also still violated!!!

Hypothesis 2:

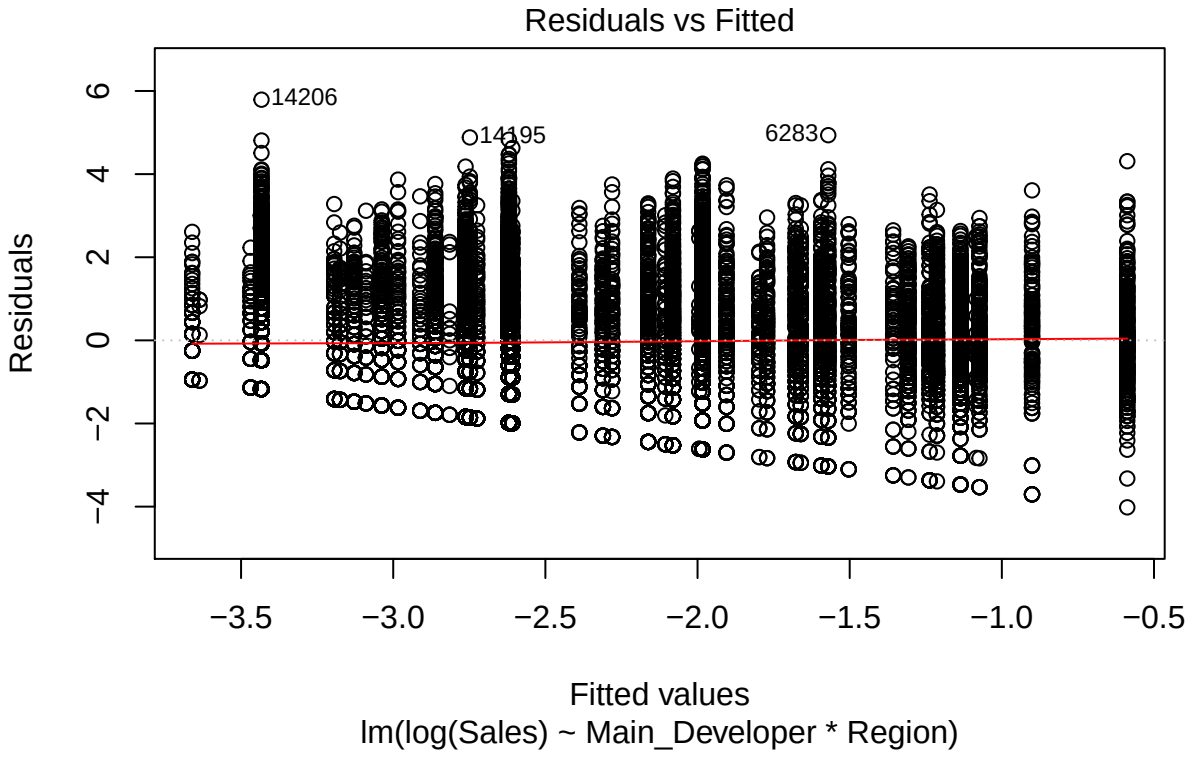
```
dev_region <- lm(log(Sales) ~ Main_Developer*Region, data = sales_nonzero)
dev_region %>% anova()
```

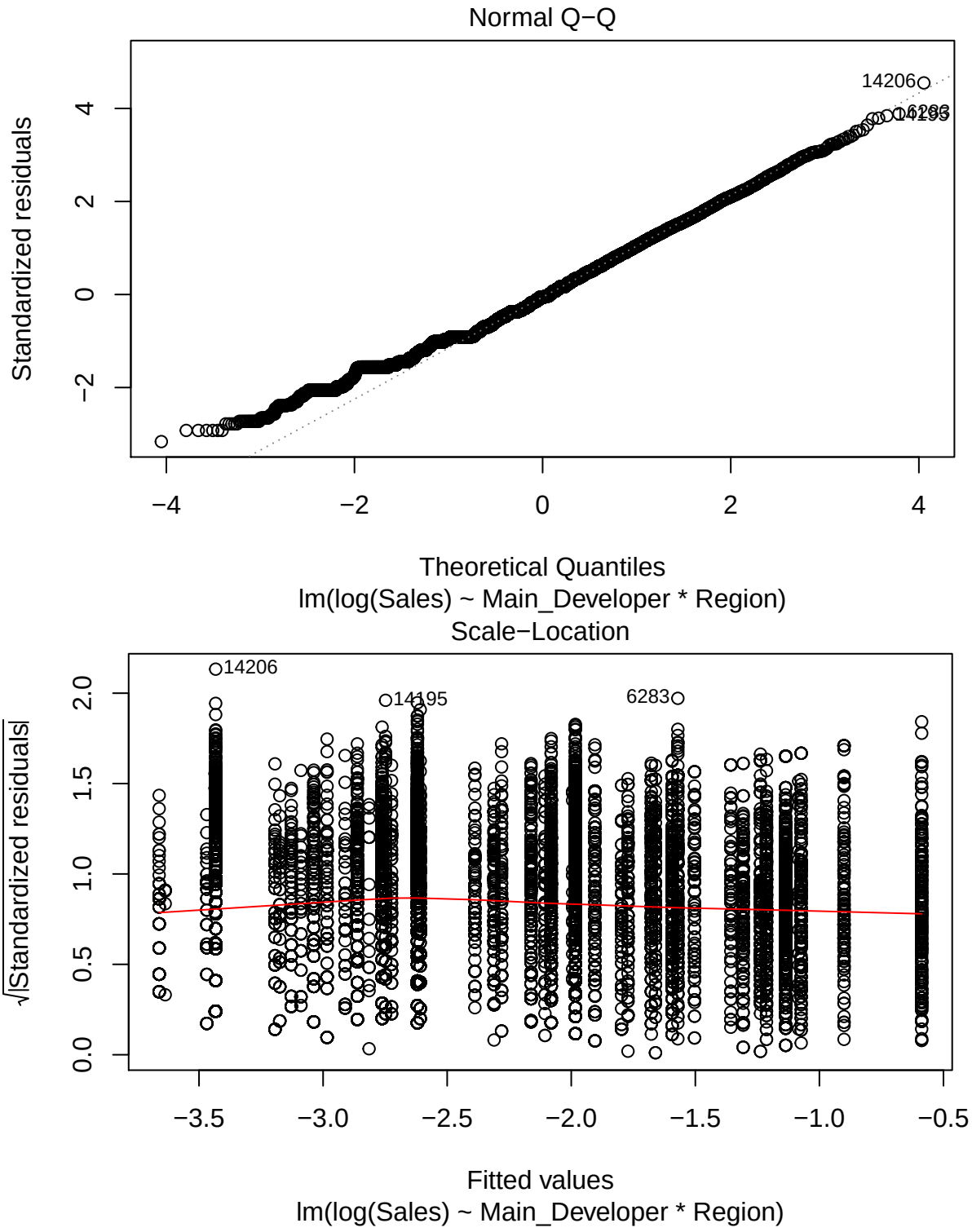
```
## Analysis of Variance Table
##
```

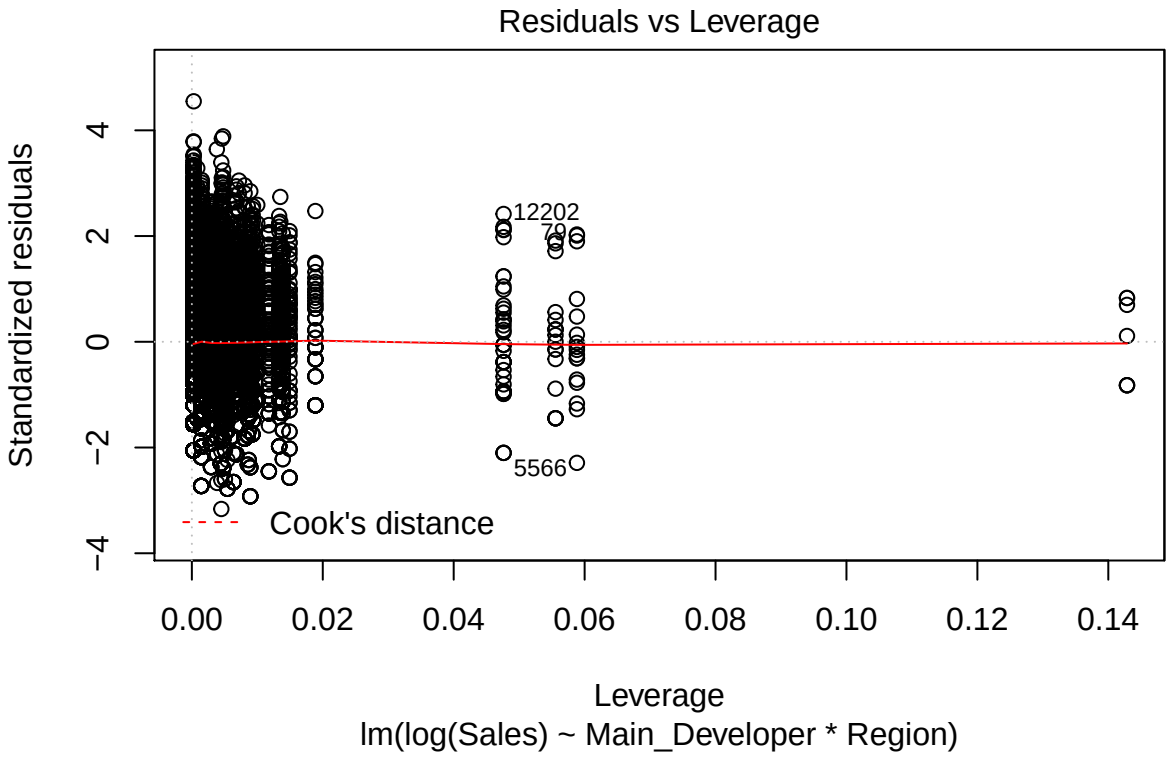
```
## Response: log(Sales)
##                            Df Sum Sq Mean Sq  F value    Pr(>F)
## Main_Developer             10   2049  204.86  126.379 < 2.2e-16 ***
## Region                      3   7056 2352.15 1451.019 < 2.2e-16 ***
## Main_Developer:Region      30    777   25.90   15.976 < 2.2e-16 ***
## Residuals               19750  32015    1.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check the assumptions:

```
plot(dev_region)
```




Normal Q−Q
Standardized residuals
14206
Theoretical Quantiles
lm(log(Sales) ~ Main_Developer * Region)
Scale−Location
√|Standardized residuals|
14206
14195
6283
Fitted values
lm(log(Sales) ~ Main_Developer * Region)



Based on the QQ plot and residual plot there isn't much deviation from the normality and homogenity assumptions. There appears to be a small pattern in residual plot, we will check if it's significant by additional tests:

```
shapiro.test(dev_region$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dev_region$residuals[sample(1:length(sales_nonzero$Platform),     5000)]
## W = 0.99172, p-value < 2.2e-16
```

We see that the normality assumption is actually violated. What about homogeneity? - It's also violated!!! There are no outliers and influential points.

```
bptest(dev_region)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  dev_region
## BP = 436.4, df = 43, p-value < 2.2e-16
```

WLS method might be a potential remedy for the assumptions violations. So try this:

```
varfunc_dev_reg <- lm(log(dev_region$residuals^2) ~ Main_Developer*Region,
                      data = sales_nonzero)
sales_nonzero$varfunc_dev_reg <- exp(varfunc_dev_reg$fitted.values)
dev_region_gls <- lm(log(Sales) ~ Main_Developer*Region,
                       weights = 1/sqrt(varfunc_dev_reg),
                       data = sales_nonzero)
```

Check normality:

```
shapiro.test(dev_region_gls$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dev_region_gls$residuals[sample(1:length(sales_nonzero$Platform),     5000)]
## W = 0.99275, p-value = 2.811e-15
```

P-value is strictly less than 0.05 – reject the null hypothesis stating the normality of residuals! NOT normal!!!! Hence WLS method didn't help to resolve the abnormality :D issue.

```
bptest(dev_region_gls)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  dev_region_gls
## BP = 436.4, df = 43, p-value < 2.2e-16
```

Homogeneity is also still violated!!!

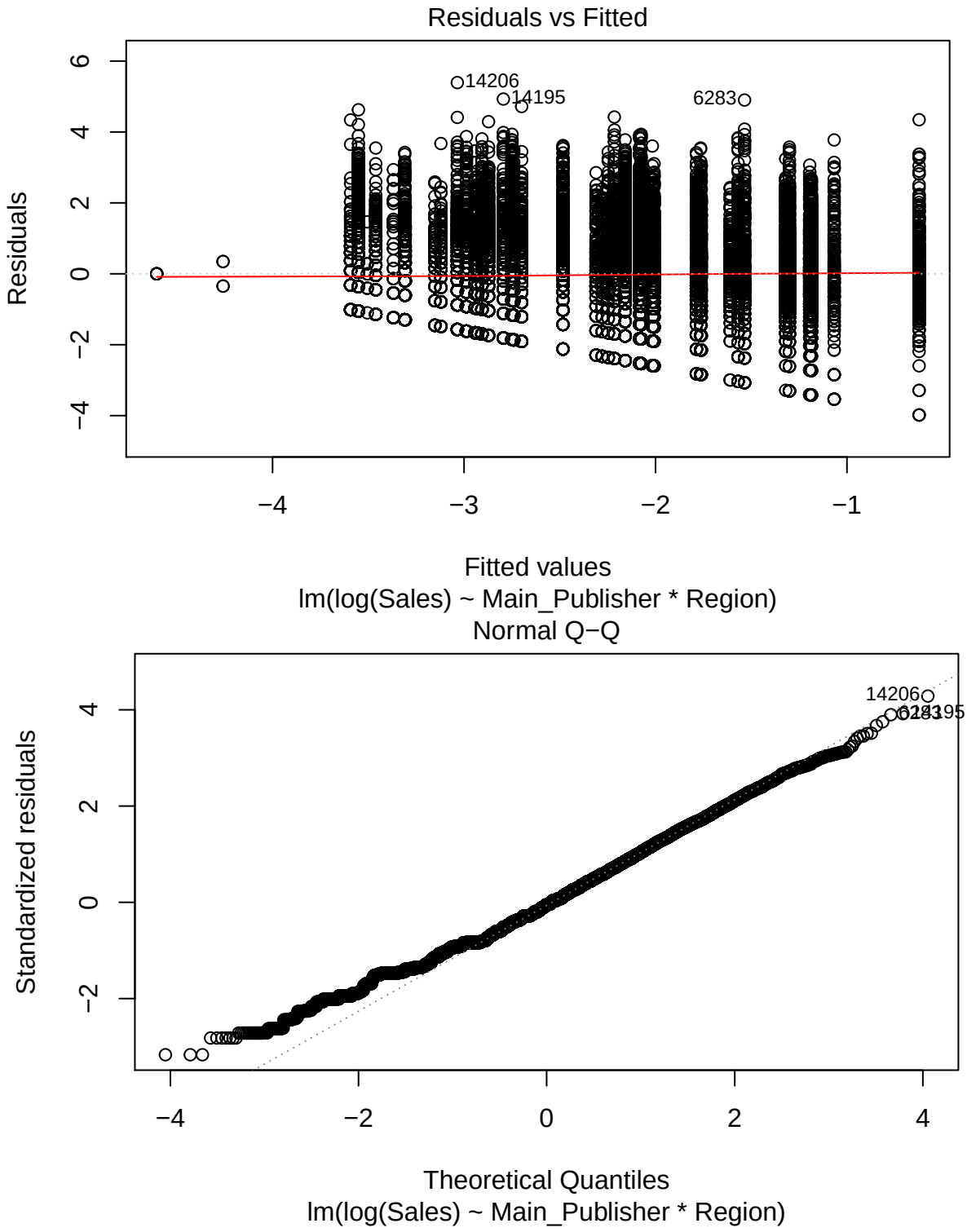
Hypothesis 3

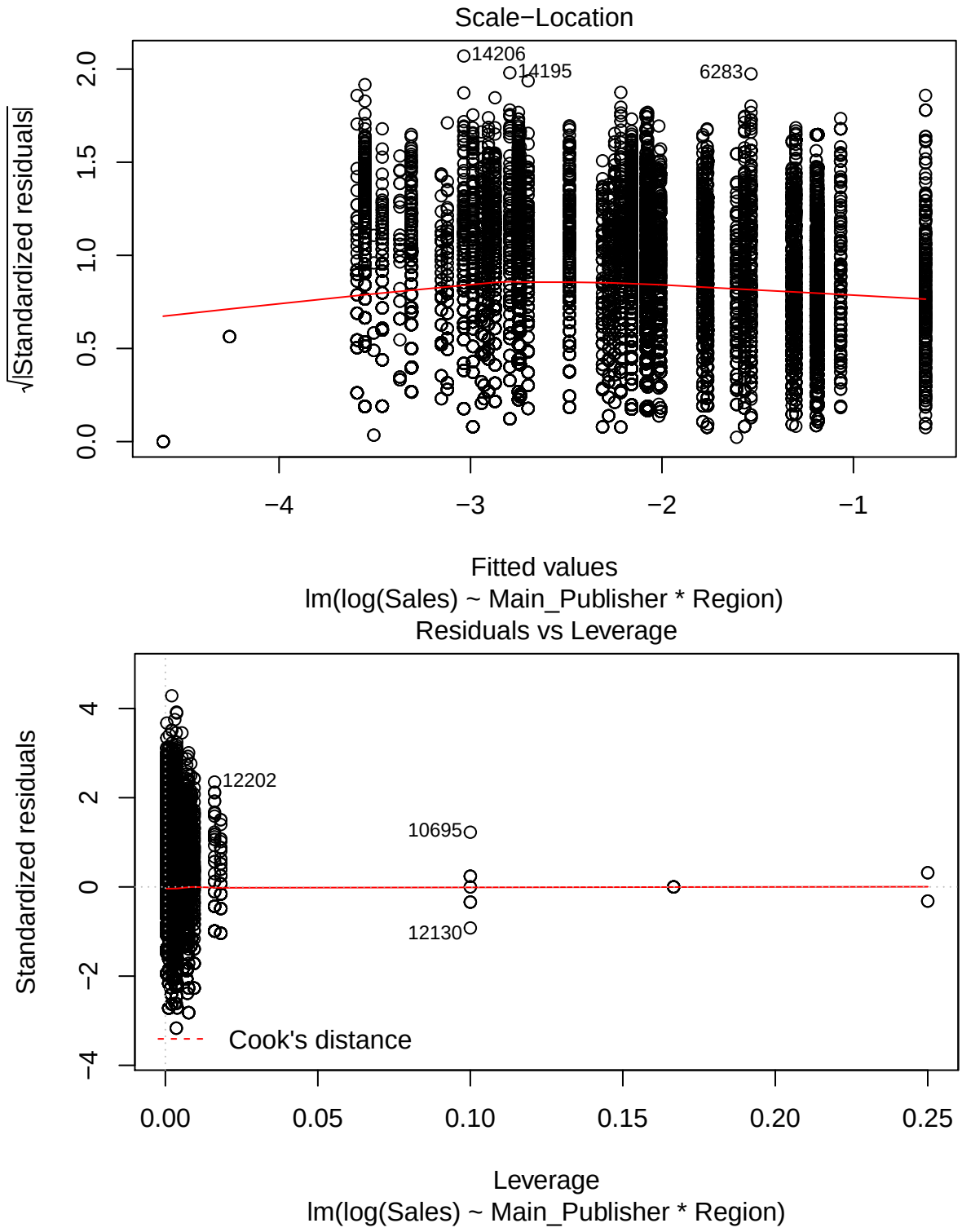
```
pub_region <- lm(log(Sales) ~ Main_Publisher*Region, data = sales_nonzero)
pub_region %>% anova()
```

```
## Analysis of Variance Table
##
## Response: log(Sales)
##                          Df  Sum Sq Mean Sq  F value    Pr(>F)
## Main_Publisher           10  2621.5  262.15  165.223 < 2.2e-16 ***
## Region                    3  7103.5 2367.84 1492.371 < 2.2e-16 ***
## Main_Publisher:Region    29   834.9   28.79   18.145 < 2.2e-16 ***
## Residuals             19751 31337.5    1.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check the assumptions:

```
plot(pub_region)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(Sales) ~ Main_Publisher * Region)
14206
14195
6283
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(Sales) ~ Main_Publisher * Region)



Based on the QQ plot and residual plot there isn't much deviation from the normality and homogenity assumptions. There appears to be a small pattern in residual plot, we will check if it's significant by additional tests:

```
shapiro.test(pub_region$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pub_region$residuals[sample(1:length(sales_nonzero$Platform),     5000)]
## W = 0.99292, p-value = 4.504e-15
```

We see that the normality assumption is actually violated. What about homogeneity? - It's also violated!!! There are no outliers and influential points.

```
bptest(pub_region)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  pub_region
## BP = 545.08, df = 42, p-value < 2.2e-16
```

WLS method might be a potential remedy for the assumptions violations. So try this:

```
varfunc_pub_reg <- lm(log(pub_region$residuals^2) ~ Main_Publisher*Region,
                      data = sales_nonzero)
sales_nonzero$varfunc_pub_reg <- exp(varfunc_pub_reg$fitted.values)
pub_region_gls <- lm(log(Sales) ~ Main_Publisher*Region,
                       weights = 1/sqrt(varfunc_pub_reg),
                       data = sales_nonzero)
```

Check normality:

```
shapiro.test(pub_region_gls$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pub_region_gls$residuals[sample(1:length(sales_nonzero$Platform),     5000)]
## W = 0.99267, p-value = 2.241e-15
```

P-value has increased but still strictly less than 0.05 – reject the null hypothesis stating the normality of residuals! NOT normal!!!! Hence WLS method didn't help to resolve the abnormality :D issue.

```
bptest(pub_region_gls)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  pub_region_gls
## BP = 545.08, df = 42, p-value < 2.2e-16
```

Homogeneity is also still violated!!!

Hypothesis 4

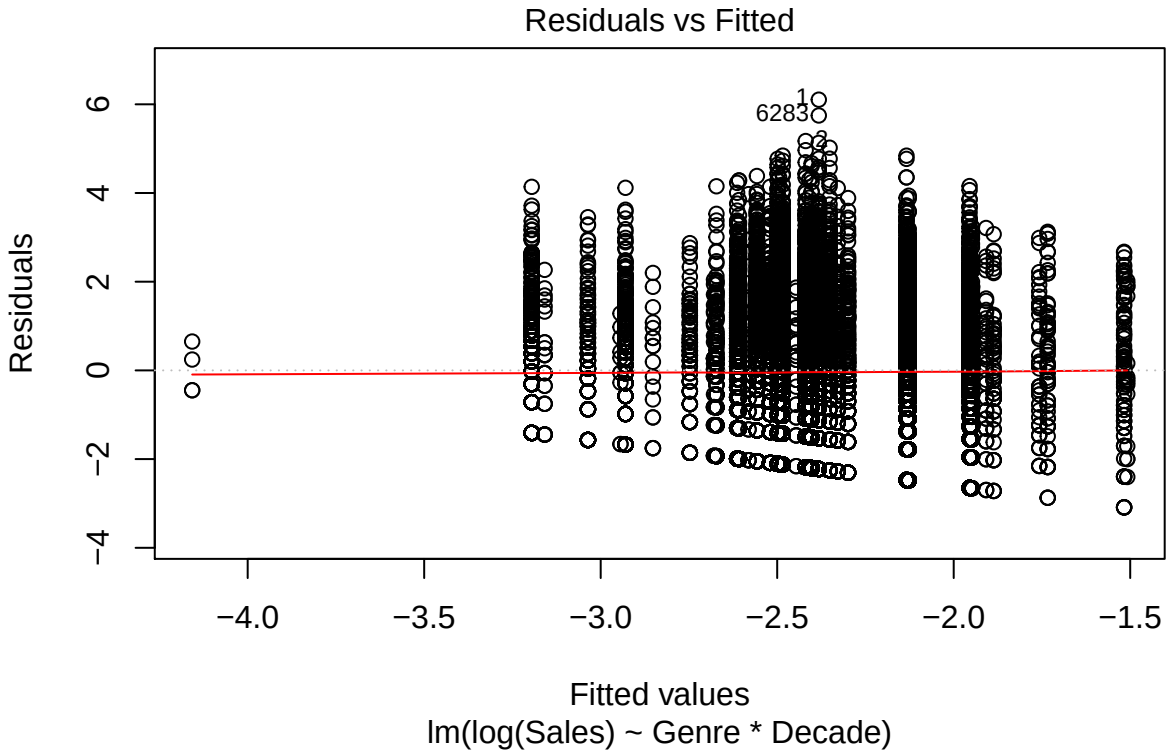
```
gen_dec <- lm(log(Sales) ~ Genre*Decade, data = sales_nonzero)
gen_dec %>% anova()
```

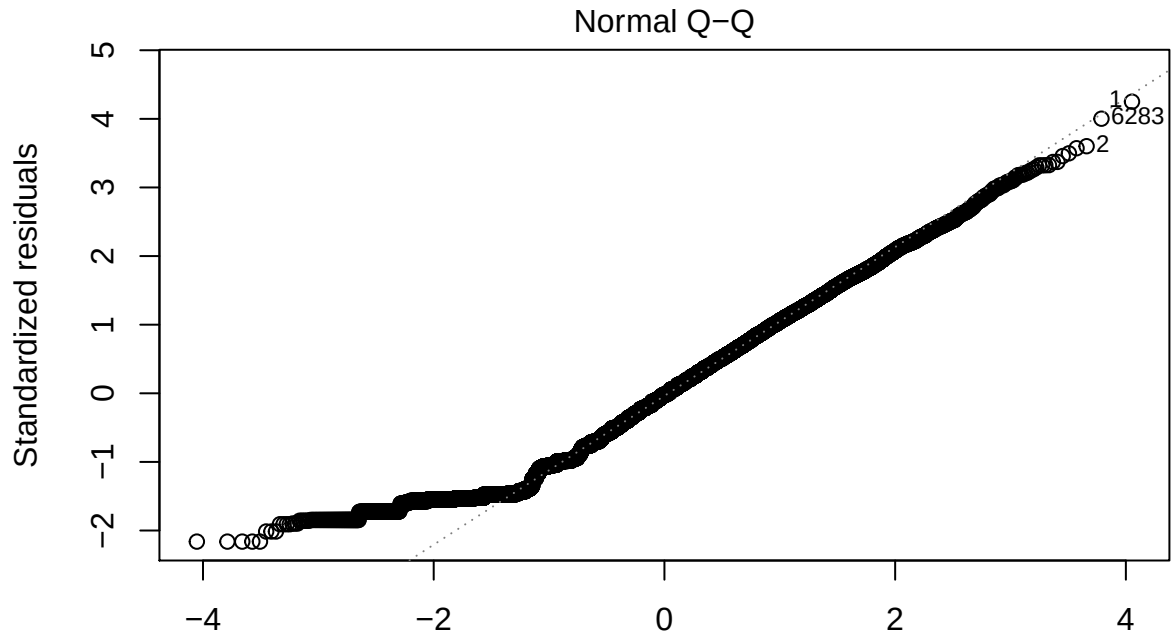
```
## Analysis of Variance Table
##
```

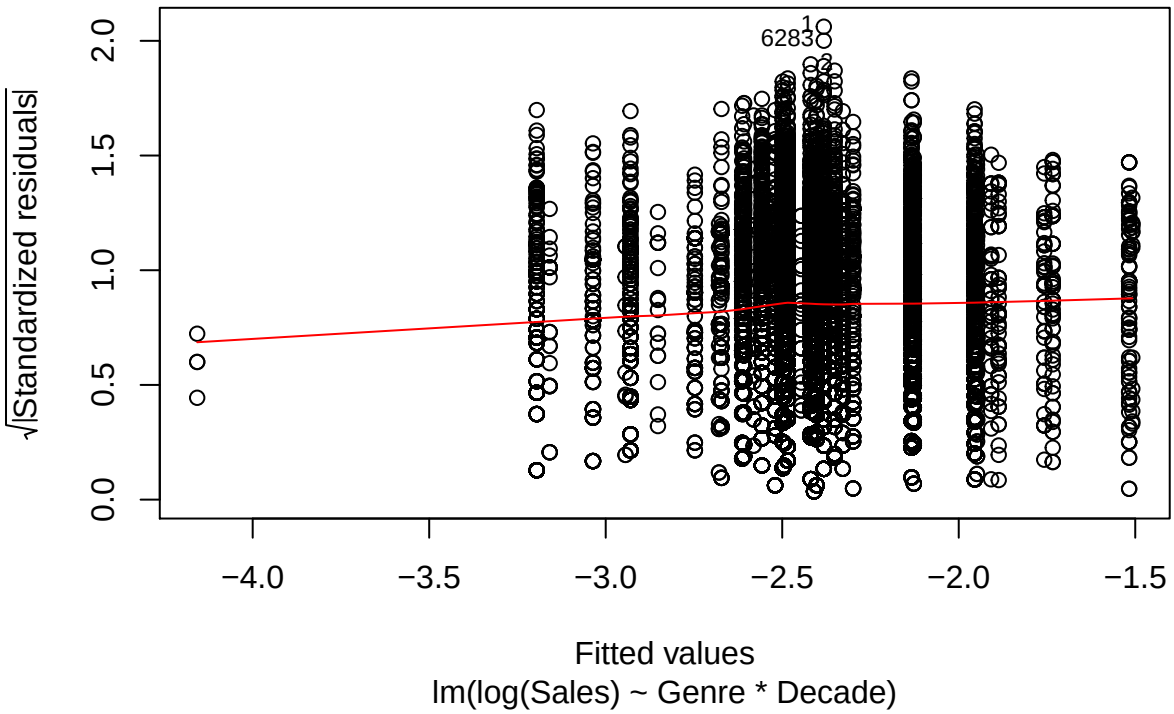
```
## Response: log(Sales)
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## Genre           11    661  60.076 29.1101 < 2.2e-16 ***
## Decade           3    191  63.800 30.9144 < 2.2e-16 ***
## Genre:Decade    22    272  12.343  5.9806 < 2.2e-16 ***
## Residuals    19757  40774   2.064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

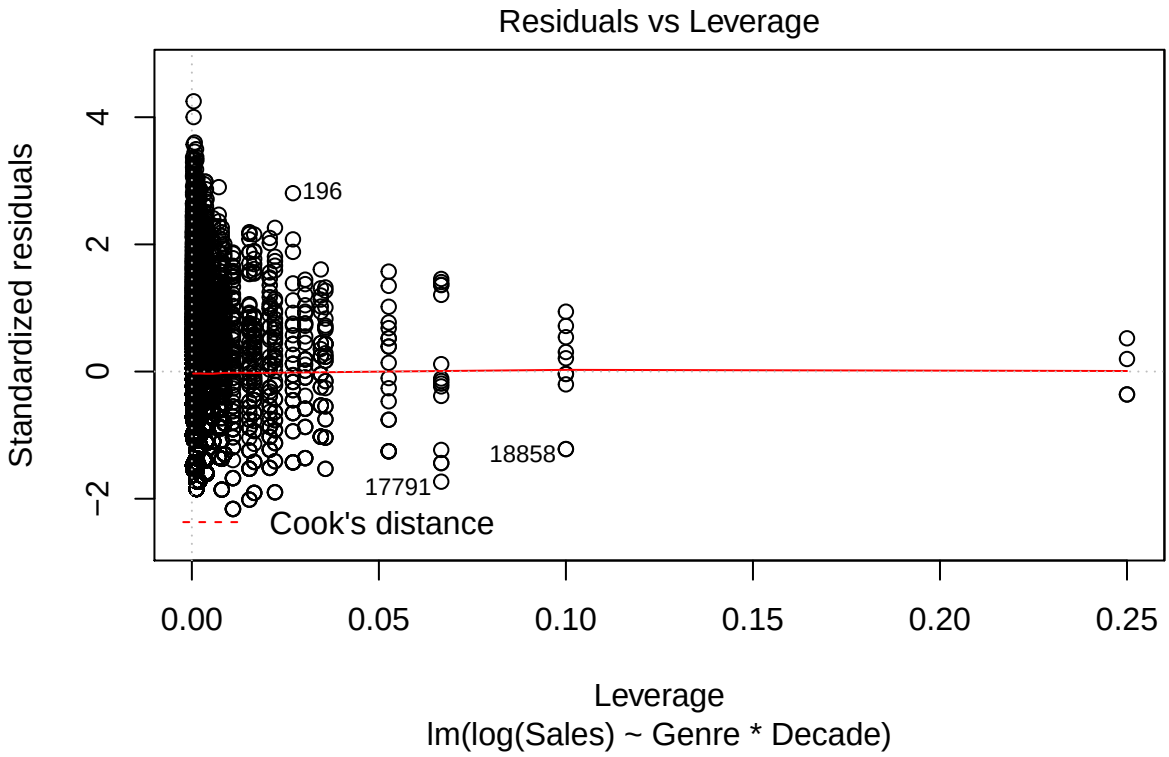
Check the assumptions:

```
plot(gen_dec)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(Sales) ~ Genre * Decade)
1
6283
2


Scale−Location
√|Standardized residuals|
Fitted values
lm(log(Sales) ~ Genre * Decade)
1
6283



Based on the QQ plot and residual plot there isn't much deviation from the normality and homogenity assumptions. There appears to be a small pattern in residual plot, we will check if it's significant by additional tests:

```
shapiro.test(gen_dec$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gen_dec$residuals[sample(1:length(sales_nonzero$Platform), 5000)]
## W = 0.98058, p-value < 2.2e-16
```

We see that the normality assumption is actually violated. What about homogeneity? - It's also violated!!! There are no outliers and influential points.

```
bptest(gen_dec)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  gen_dec
## BP = 168.23, df = 36, p-value < 2.2e-16
```

WLS method might be a potential remedy for the assumptions violations. So try this:

```
varfunc_gen_dec <- lm(log(gen_dec$residuals^2) ~ Genre*Decade,
                      data = sales_nonzero)
sales_nonzero$varfunc_gen_dec <- exp(varfunc_gen_dec$fitted.values)
gen_dec_gls <- lm(log(Sales) ~ Genre*Decade,
                       weights = 1/sqrt(varfunc_gen_dec),
                       data = sales_nonzero)
```

Check normality:

```
shapiro.test(gen_dec_gls$residuals
             [sample(1:length(sales_nonzero$Platform), 5000)])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gen_dec_gls$residuals[sample(1:length(sales_nonzero$Platform),     5000)]
## W = 0.979, p-value < 2.2e-16
```

P-value is strictly less than 0.05 – reject the null hypothesis stating the normality of residuals! NOT normal!!!! Hence WLS method didn't help to resolve the abnormality :D issue.

```
bptest(gen_dec_gls)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  gen_dec_gls
## BP = 168.23, df = 36, p-value < 2.2e-16
```

Homogeneity is also still violated!!!