Micro Proj 1

Sierra Rossman

2025-06-07

Research question: To what extent do country and year predict child RCV1 (rubella) vaccine coverage in Eastern Europe & Central Asia from 1980 to 2023?

Outcome variable: mean Predictors: year_id, location_name filter by: vaccine_name = rcv1, location_id = “Russia”, “Ukraine”, “Belarus”, “Georgia”, “Armenia”, “Kazakhstan”, “Kyrgyzstan”, “Uzbekistan”, “Moldova”, or “Azerbaijan”.

#Data Cleaning/Preparation

```
#Importing the dataset
data <- read.csv("IHME_GBD_2023_VACC_1980_2030_ESTIMATES_Y2025M04D25.CSV")
#Filtering the dataset to my research question
rcv1_data <- data %>% filter(vaccine_name == "RCV1")
eecca_countries <- c(
  "Russia", "Ukraine", "Belarus", "Georgia", "Armenia",
  "Kazakhstan", "Kyrgyzstan", "Uzbekistan", "Moldova", "Azerbaijan"
)
rcv1_data <- rcv1_data %>% filter(location_name %in% eecca_countries)
#Looking at the dataset to ensure everything worked as intended
head(rcv1_data)
```

```
##   vaccine_name location_name year_id mean
## 1         RCV1       Armenia    1980    0
## 2         RCV1       Armenia    1981    0
## 3         RCV1       Armenia    1982    0
## 4         RCV1       Armenia    1983    0
## 5         RCV1       Armenia    1984    0
## 6         RCV1       Armenia    1985    0
```

Cleaning:

```
#Getting the total number of missing values
sum(is.na(rcv1_data)) #There were none so handling wasn't necessary
```

```
## [1] 0
```

```
#Reformatting vaccine_name and location_name
rcv1_data$location_name <- as.factor(rcv1_data$location_name)
#Getting summary statistics
summary(rcv1_data)
```

```
##  vaccine_name          location_name    year_id          mean
##  Length:384         Armenia   :48    Min.   :1980   Min.   :0.0000
##  Class :character   Azerbaijan:48    1st Qu.:1992   1st Qu.:0.0000
##  Mode  :character   Belarus   :48    Median :2004   Median :0.7395
##                     Georgia   :48    Mean   :2003   Mean   :0.4799
##                     Kazakhstan:48    3rd Qu.:2015   3rd Qu.:0.9273
##                     Kyrgyzstan:48    Max.   :2023   Max.   :0.9900
##                     (Other)   :96
```

```
#Getting the standard deviation for the mean variable which is the
#only continuous variable that makes sense to get a standard deviation stat for.
sd(rcv1_data$mean)
```

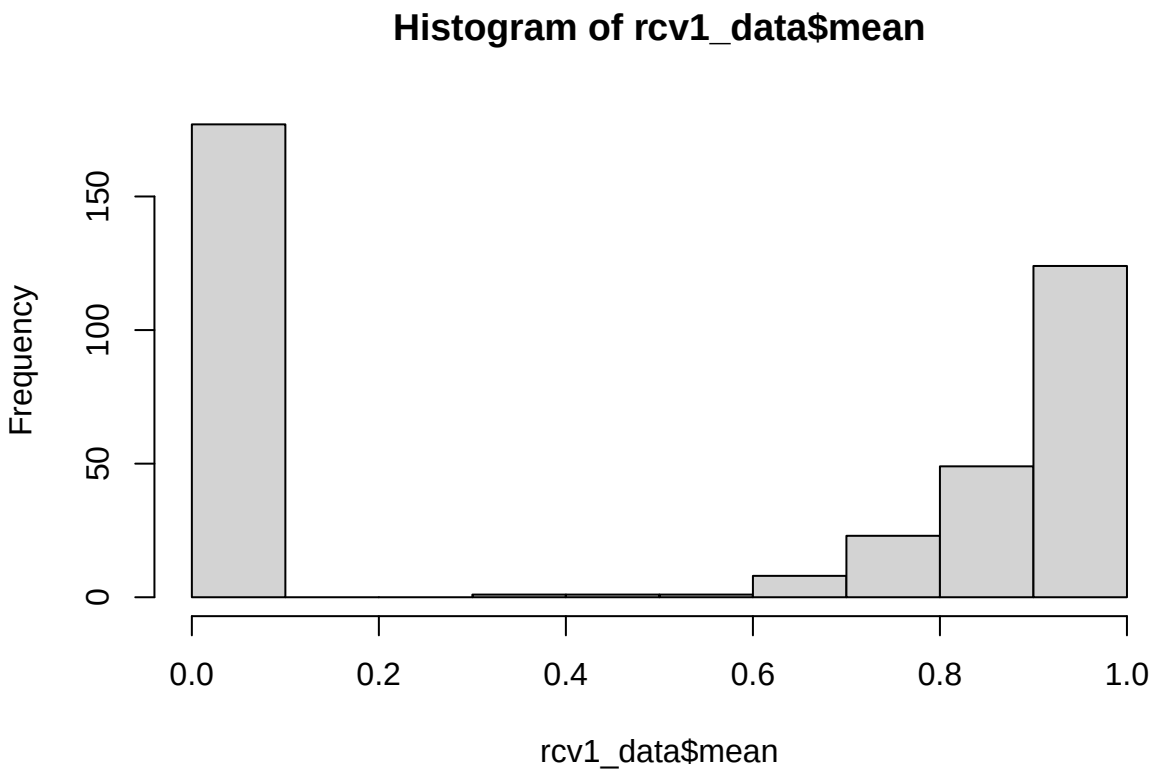
```
## [1] 0.4499626
```

```
#Showing the result of the data preparation steps
head(rcv1_data)
```

```
##   vaccine_name location_name year_id mean
## 1         RCV1       Armenia    1980    0
## 2         RCV1       Armenia    1981    0
## 3         RCV1       Armenia    1982    0
## 4         RCV1       Armenia    1983    0
## 5         RCV1       Armenia    1984    0
## 6         RCV1       Armenia    1985    0
```

#Multiple Linear Regression Null Hypotheses: There is no relationship between year and RCV1 vaccine coverage There is no difference in RCV1 vaccine coverage between countries, holding year constant Alternative Hypotheses: There is a relationship between year and RCV1 vaccine coverage At least one country differs significantly in coverage compared to the reference country

```
#First looking at the distribution of the mean variable
#Noted that there are a lot of 0's which may cause problems with the assumptions
hist(rcv1_data$mean)
```

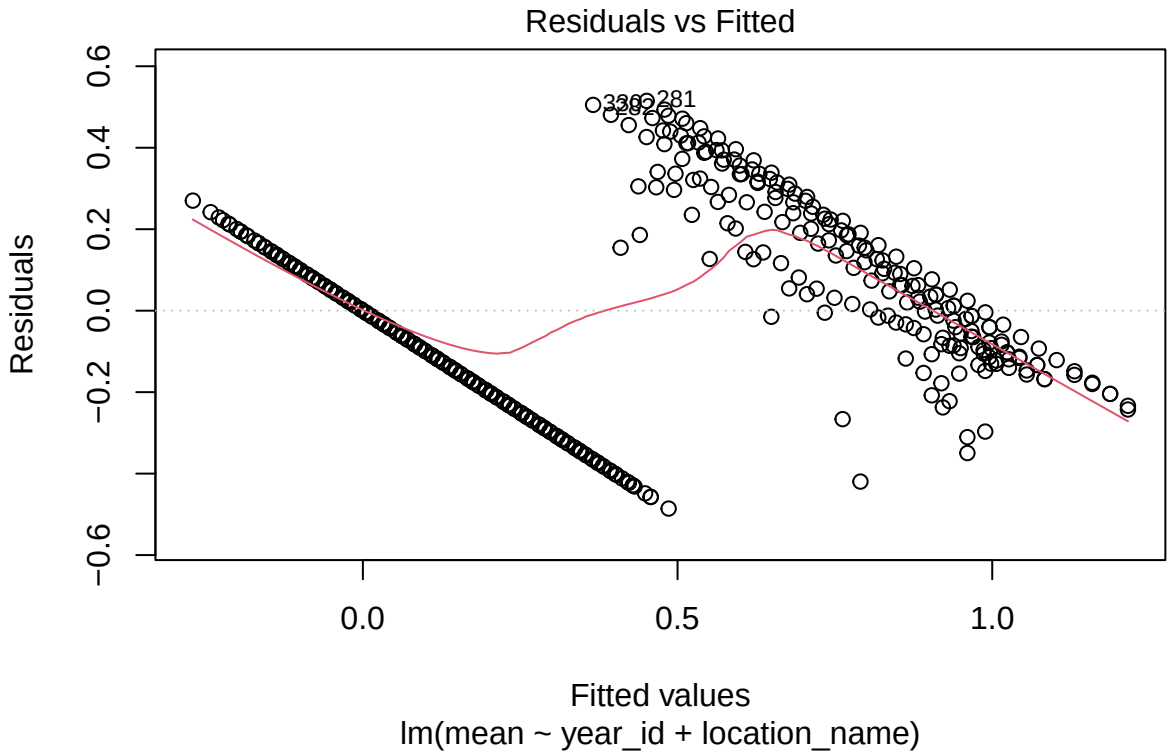



```r
#Fitting the first model
model <- lm(mean ~ year_id + location_name, data = rcv1_data)
#Outputting the results from the model
summary(model)
```

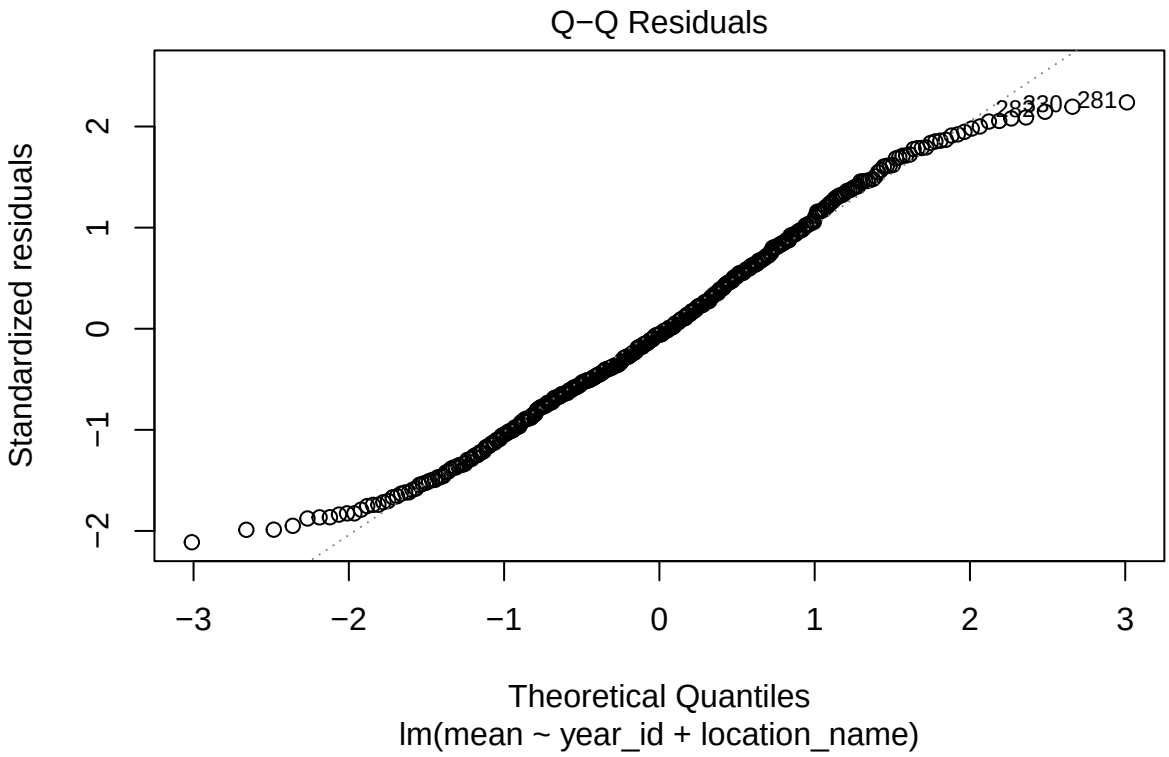
```
##
## Call:
## lm(formula = mean ~ year_id + location_name, data = rcv1_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48581 -0.15710 -0.01313  0.16002  0.51497
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -5.622e+01  1.779e+00 -31.593  < 2e-16 ***
## year_id                   2.832e-02  8.881e-04  31.886  < 2e-16 ***
## location_nameAzerbaijan -6.498e-02  4.746e-02  -1.369  0.17177
## location_nameBelarus     1.442e-01  4.746e-02   3.038  0.00255 **
## location_nameGeorgia    -1.239e-01  4.746e-02  -2.611  0.00940 **
## location_nameKazakhstan -4.790e-02  4.746e-02  -1.009  0.31353
## location_nameKyrgyzstan  1.165e-02  4.746e-02   0.245  0.80629
## location_nameUkraine    -8.260e-02  4.746e-02  -1.741  0.08259 .
## location_nameUzbekistan -7.590e-02  4.746e-02  -1.599  0.11062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2325 on 375 degrees of freedom
## Multiple R-squared:  0.7386, Adjusted R-squared:  0.733
## F-statistic: 132.4 on 8 and 375 DF,  p-value: < 2.2e-16
```

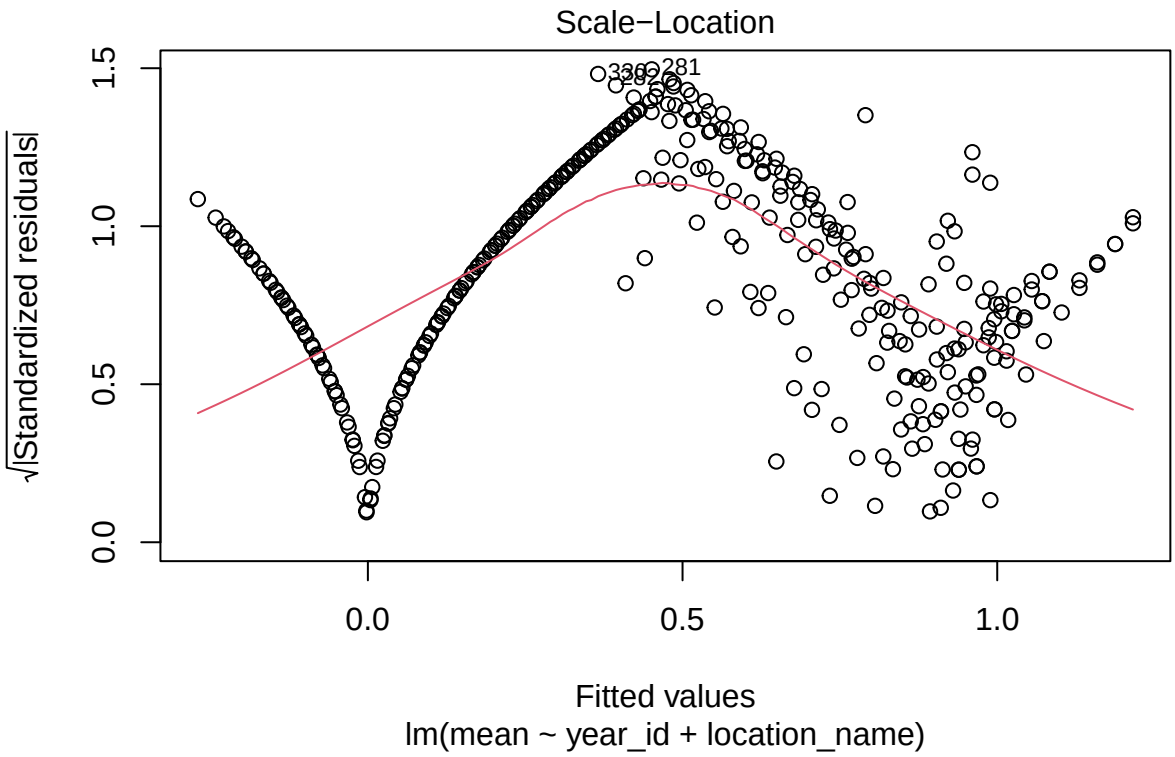
```r
#Checking the assumptions
#Plot 1: Residuals vs Fitted
plot(model, which = 1)
```



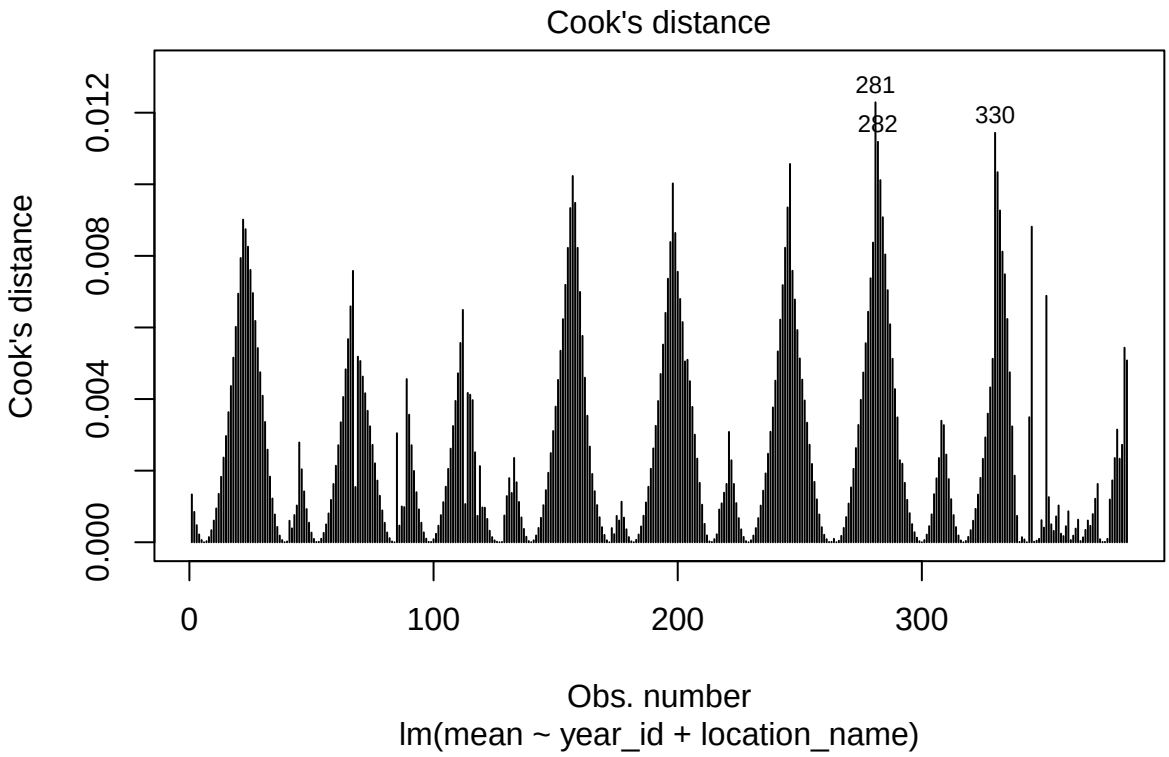

```r
#Plot 2: Normal Q-Q Plot
plot(model, which = 2)
```




```
#Plot 3: Scale-Location Plot Spread vs fitted homoscedasticity or equal variance
plot(model, which = 3)
```




```
#Plot 4: Residuals vs leverage potential influential outliers
plot(model, which = 4)
```




```
#Aside from the Q-Q plot, all of the assumptions fail
```

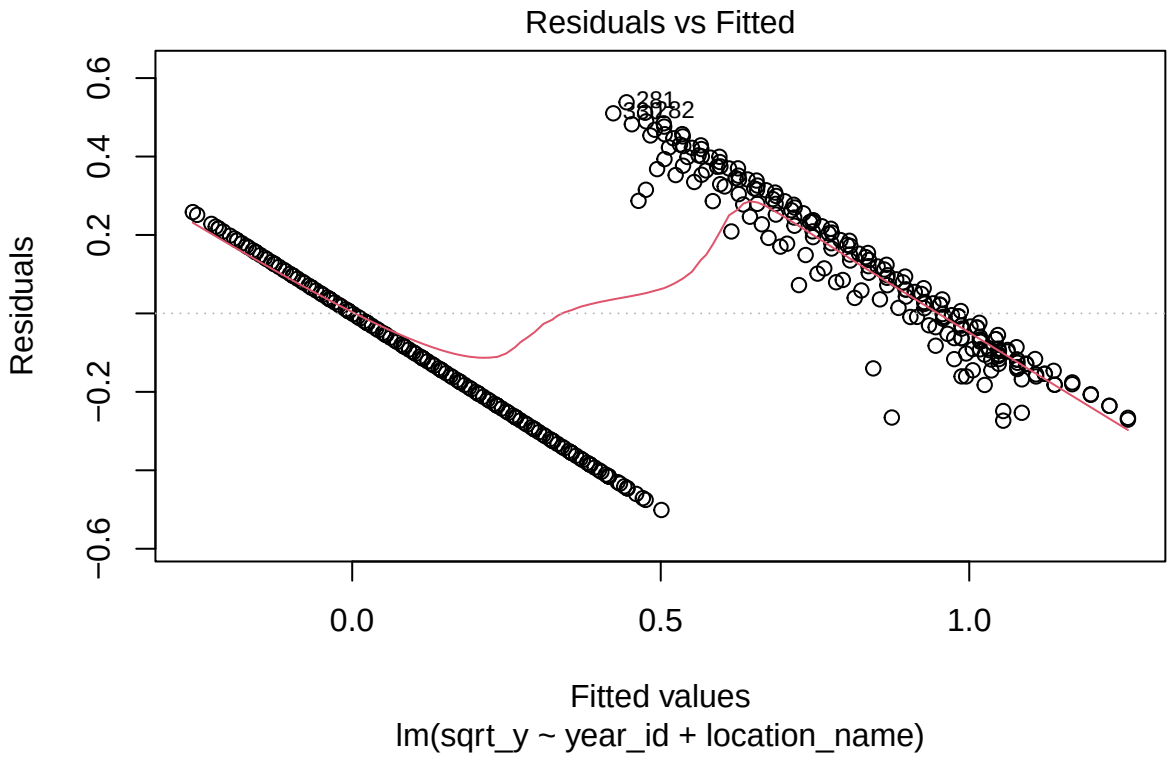
```
#Attempt to transform the mean variable to satisfy the assumptions
sqrt_y <- sqrt(rcv1_data$mean)
#Fitting the second model based on the transformed response
model2 <- lm(sqrt_y ~ year_id + location_name, data = rcv1_data)
```

```
#Seeing how the new model performed - increase in R-squared noted
summary(model2)
```

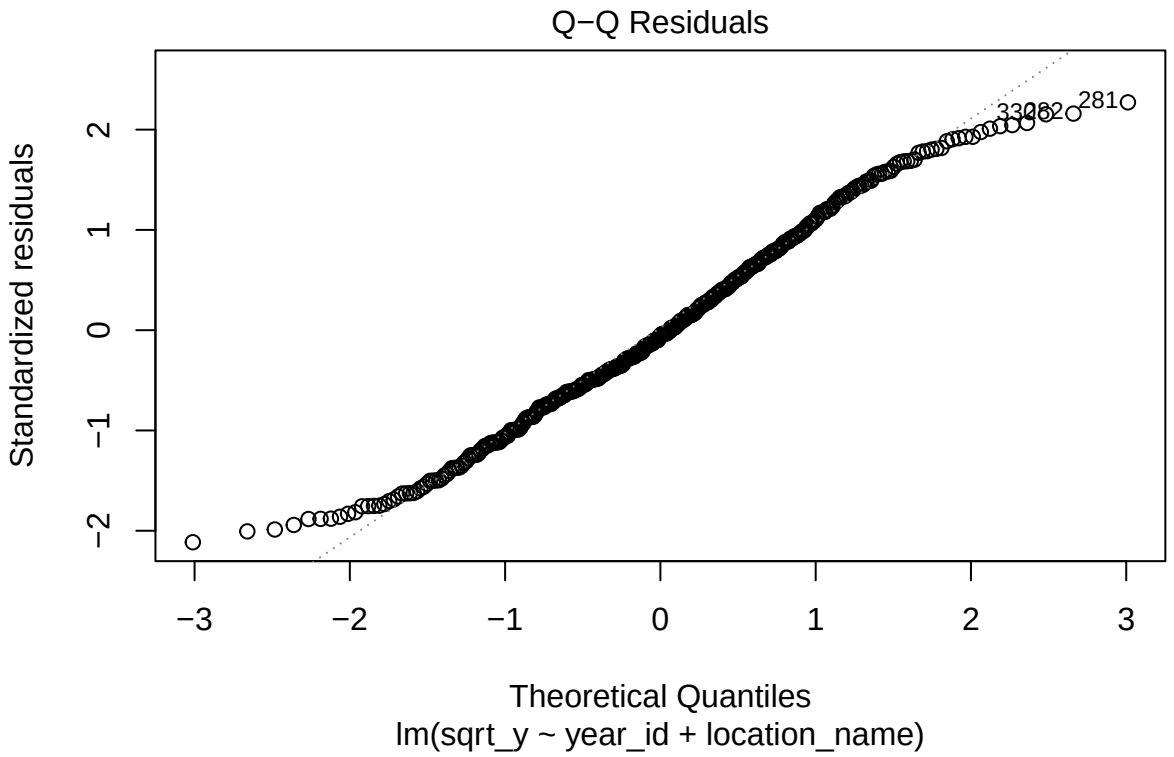
```
##
## Call:
## lm(formula = sqrt_y ~ year_id + location_name, data = rcv1_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50112 -0.16129 -0.01306  0.17181  0.53825
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -5.976e+01  1.833e+00 -32.604   <2e-16 ***
## year_id                   3.010e-02  9.149e-04  32.897   <2e-16 ***
## location_nameAzerbaijan -4.449e-02  4.889e-02  -0.910   0.3634
## location_nameBelarus     1.348e-01  4.889e-02   2.758   0.0061 **
## location_nameGeorgia    -8.652e-02  4.889e-02  -1.770   0.0776 .
```

```
## location_nameKazakhstan -4.493e-02  4.889e-02  -0.919   0.3586
## location_nameKyrgyzstan  1.607e-02  4.889e-02   0.329   0.7426
## location_nameUkraine    -3.715e-02  4.889e-02  -0.760   0.4478
## location_nameUzbekistan -7.951e-02  4.889e-02  -1.626   0.1047
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2395 on 375 degrees of freedom
## Multiple R-squared:  0.7477, Adjusted R-squared:  0.7424
## F-statistic: 138.9 on 8 and 375 DF,  p-value: < 2.2e-16
```

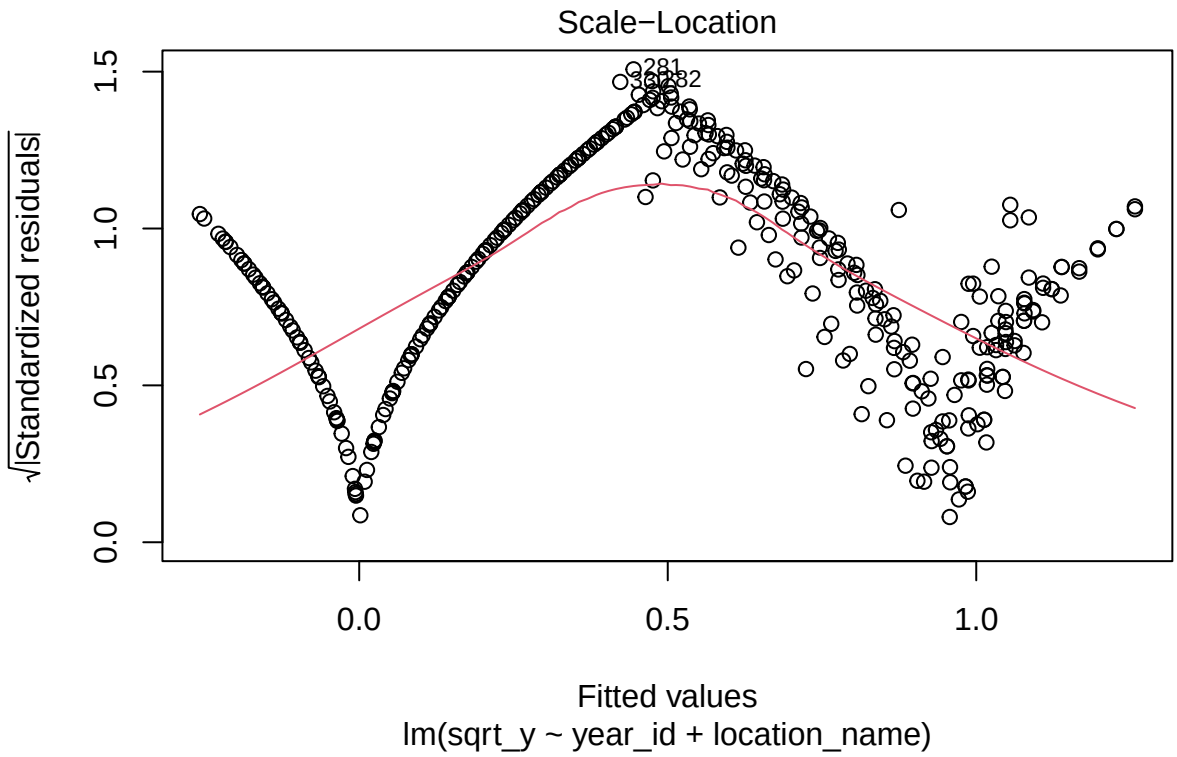
```r
#Plot 1: Residuals vs Fitted
plot(model2, which = 1)
```



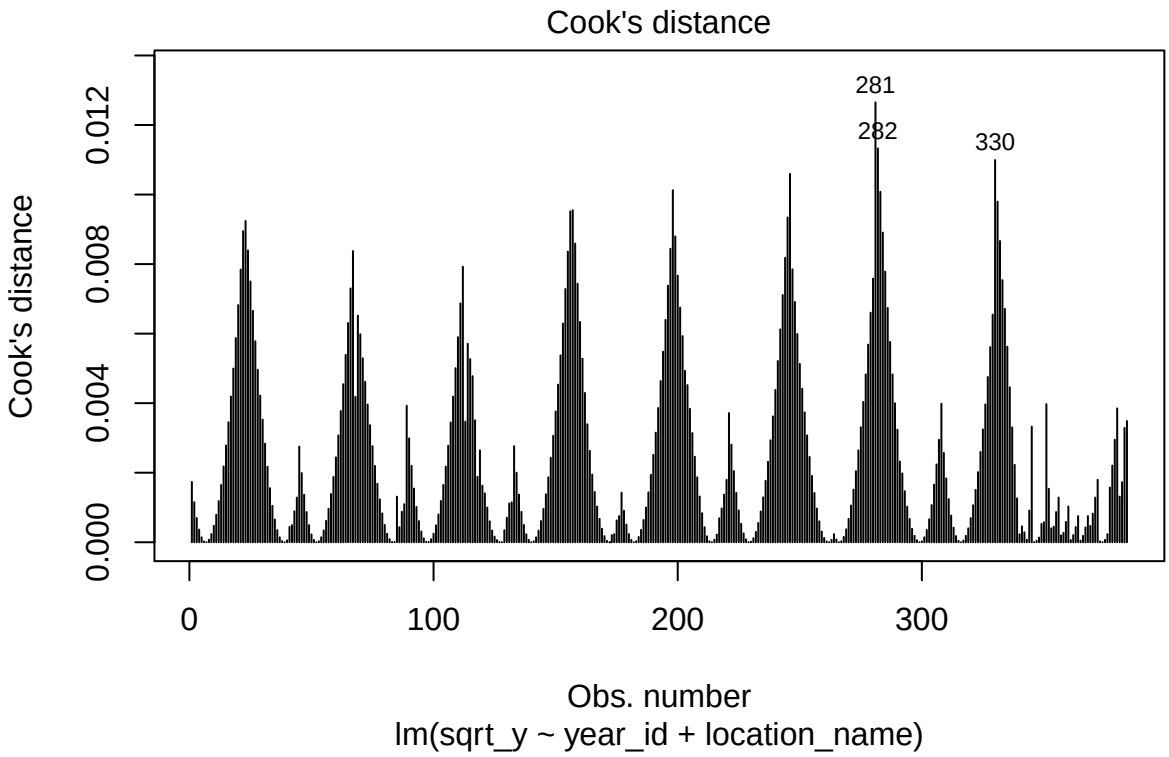

```r
#Plot 2: Normal Q-Q Plot
plot(model2, which = 2)
```




```
#Plot 3: Scale-Location Plot Spread vs fitted homoscedasticity or equal variance
plot(model2, which = 3)
```




```
#Plot 4: Residuals vs leverage potential influenctial outliers
plot(model2, which = 4)
```




```
#Again, aside from the Q-Q plot, the assumptions fail so I will infer/conclude with caution.
```

#Inferences

Mean is the mean vaccine coverage in children, specifically filtered for the Eastern European and Central Asia regions and also filtered for the RCV1 vaccine. Year is the year the data was recorded, from 1980 to 2023. Country is the country the data was recorded from. An increase in year is correlated with an increase of 0.0301, holding country constant, with a p-value of <2e-16. Belarus has a statistically significant (p=0.0061) higher RCV1 vaccine coverage than the baseline country (0.1348). Considering a p-value threshold of 0.05, those are the two statistically significant results. Increasing the p-value threshold < 0.1, Georgia has a statistically significant (p=0.0776) lower RCV1 vaccine coverage than the baseline country (-0.08652). R-squared is equal to 0.7477 which tells us 74.77% of the variance is explained by year and country which indicates a strong overall fit. The f-statistic is < 2.2e-16, proving the model effective and statistically significant. We can therefore reject our null hypotheses with alpha < 0.05.

#Conclusions

Mean RCV1 vaccine coverage in children has increased within the Eastern European and Central Asia regions over the years. Belarus specifically has seen a significantly increased mean coverage of the RCV1 vaccine compared to the reference country. Alternatively, Georgia had a significantly lower mean coverage of the RCV1 vaccine compared to the reference country. The overall increase of RCV1 vaccine coverage in children is encouraging for public health efforts, but more research can be done in Georgia to understand why the coverage is lower there.