Project

DSK

3/10/2022

Packages

```
library(dplyr)
```

Data

```
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 4.1.3
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   1.4.0
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

```
data <- read.csv("C:/Users/14193/Downloads/Life Expectancy Data.csv") %>%
         filter(Year==2014)
dat=data[,-c(1,2)]
dat$Status=as.factor(dat$Status)
levels(dat$Status)=c("no","yes")
```

Partitioning The Data

```
RNGkind(sample.kind = "Rounding")
```

```
## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
set.seed(0)

partition.2 <- function(data, prop.train){
  # select a random sample of size = prop.train % of total records
  selected <- sample(1:nrow(data), round(nrow(data)*prop.train), replace = FALSE)
  # create training data which has prop.train % of total records
  data.train <- data[selected,]
  # create validation data
  rest <- setdiff(1:nrow(data), selected)
  data.test <- data[rest,]
  return(list(data.train=data.train, data.test=data.test))
}

partitioned=partition.2(dat,0.7)
training.data=partitioned$data.train
test.data=partitioned$data.test
```

Imputation on training data and using the attributes of training data on test data

```
sapply(training.data, anyNA)
```

```
##                          Status                 Life.expectancy
##                           FALSE                           FALSE
##                 Adult.Mortality                   infant.deaths
##                           FALSE                           FALSE
##                         Alcohol          percentage.expenditure
##                           FALSE                           FALSE
##                     Hepatitis.B                         Measles
##                            TRUE                           FALSE
##                             BMI               under.five.deaths
##                            TRUE                           FALSE
##                           Polio               Total.expenditure
##                           FALSE                            TRUE
##                      Diphtheria                        HIV.AIDS
##                           FALSE                           FALSE
##                             GDP                      Population
##                            TRUE                            TRUE
##            thinness..1.19.years              thinness.5.9.years
##                            TRUE                            TRUE
## Income.composition.of.resources                       Schooling
##                            TRUE                            TRUE
```

```
# BMI,Total.expenditure,GDP,thinness.5.9,Hepatities,Population,Income composition,thinnes.1.19,schoolin
```

```r
sapply(test.data, anyNA)
```

```
##                           Status                 Life.expectancy
##                            FALSE                           FALSE
##                  Adult.Mortality                   infant.deaths
##                            FALSE                           FALSE
##                          Alcohol          percentage.expenditure
##                             TRUE                           FALSE
##                      Hepatitis.B                         Measles
##                             TRUE                           FALSE
##                              BMI               under.five.deaths
##                             TRUE                           FALSE
##                            Polio               Total.expenditure
##                            FALSE                            TRUE
##                       Diphtheria                        HIV.AIDS
##                            FALSE                           FALSE
##                              GDP                      Population
##                             TRUE                            TRUE
##             thinness..1.19.years              thinness.5.9.years
##                             TRUE                            TRUE
## Income.composition.of.resources                       Schooling
##                             TRUE                            TRUE
```

```r
# HepatitisB,BMI, TOTAL EXPENDITURE,GDP,thinness.5.9,Population,Income composition,Alcohol,thinnnes.1.1

# We will first replace the missing values with the median values of the respective columns.


med.BMI.train <- median(training.data$BMI, na.rm = TRUE)
training.data$BMI[is.na(training.data$BMI)] <- med.BMI.train
test.data$BMI[is.na(test.data$BMI)] <- med.BMI.train


med.Hepatitis.train <- median(training.data$Hepatitis.B, na.rm = TRUE)
training.data$Hepatitis.B[is.na(training.data$Hepatitis.B)] <- med.Hepatitis.train
test.data$Hepatitis.B[is.na(test.data$Hepatitis.B)] <- med.Hepatitis.train

med.total_expenditure.train <- median(training.data$Total.expenditure, na.rm = TRUE)
training.data$Total.expenditure[is.na(training.data$Total.expenditure)] <- med.total_expenditure.train
test.data$Total.expenditure[is.na(test.data$Total.expenditure)] <- med.total_expenditure.train


med.GDP.train <- median(training.data$GDP, na.rm = TRUE)
training.data$GDP[is.na(training.data$GDP)] <- med.GDP.train
test.data$GDP[is.na(test.data$GDP)] <- med.GDP.train

med.thinnes1_19.train <- median(training.data$thinness..1.19.years, na.rm = TRUE)
training.data$thinness..1.19.years[is.na(training.data$thinness..1.19.years)] <- med.thinnes1_19.train
test.data$thinness..1.19.years[is.na(test.data$thinness..1.19.years)] <-  med.thinnes1_19.train
```

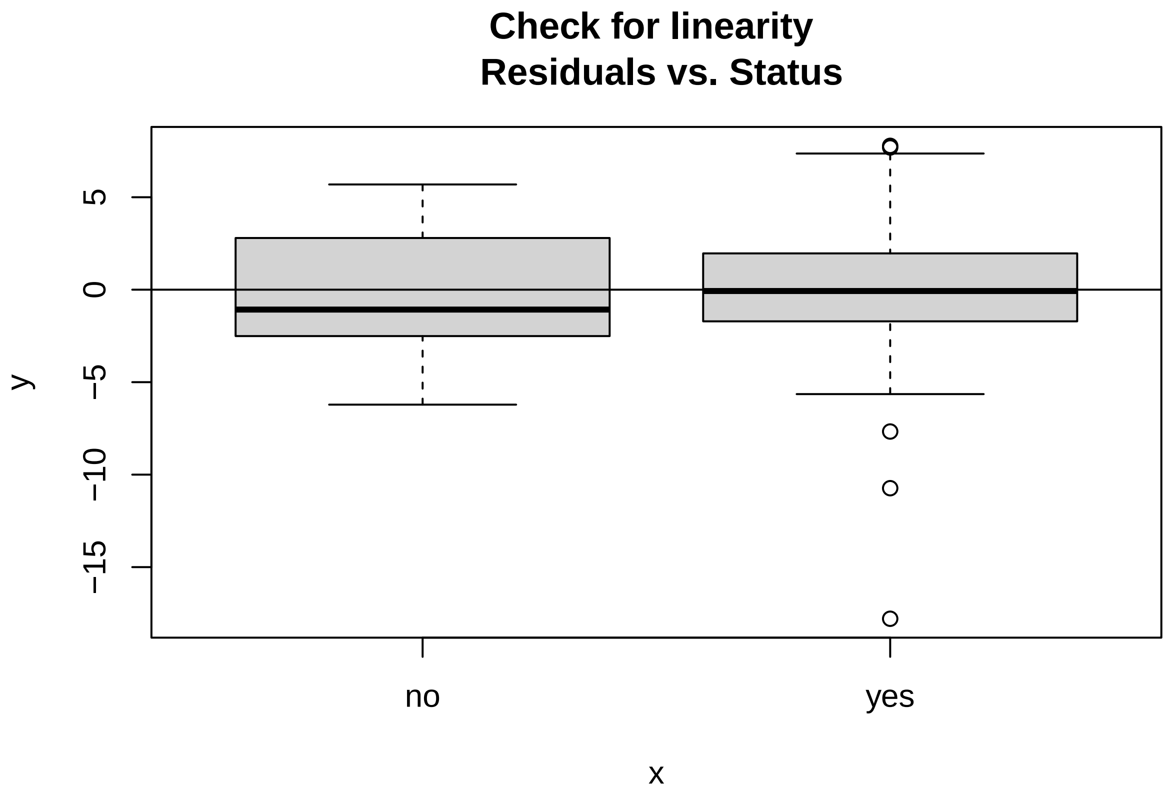
```r
med.population.train <- median(training.data$Population, na.rm = TRUE)
training.data$Population[is.na(training.data$Population)] <- med.population.train
test.data$Population[is.na(test.data$Population)] <- med.population.train


med.income.train <- median(training.data$Income.composition.of.resources, na.rm = TRUE)
training.data$Income.composition.of.resources[is.na(training.data$Income.composition.of.resources)] <- m
test.data$Income.composition.of.resources[is.na(test.data$Income.composition.of.resources)] <- med.incom

med.under5_deaths.train <- median(training.data$under.five.deaths, na.rm = TRUE)
training.data$under.five.deaths[is.na(training.data$under.five.deaths)] <- med.under5_deaths.train
test.data$under.five.deaths[is.na(test.data$under.five.deaths)] <- med.under5_deaths.train


med.alcohol.train <- median(training.data$Alcohol, na.rm = TRUE)
training.data$Alcohol[is.na(training.data$Alcohol)] <- med.alcohol.train
test.data$Alcohol[is.na(test.data$Alcohol)] <- med.alcohol.train



med.thinnes5_9.train <- median(training.data$thinness.5.9.years, na.rm = TRUE)
training.data$thinness.5.9.years[is.na(training.data$thinness.5.9.years)] <- med.thinnes5_9.train
test.data$thinness.5.9.years[is.na(test.data$thinness.5.9.years)] <- med.thinnes5_9.train


med.schooling.train <- median(training.data$Schooling, na.rm = TRUE)
training.data$Schooling[is.na(training.data$Schooling)] <- med.schooling.train
test.data$Schooling[is.na(test.data$Schooling)] <- med.schooling.train
```

Fitting a model on the training data

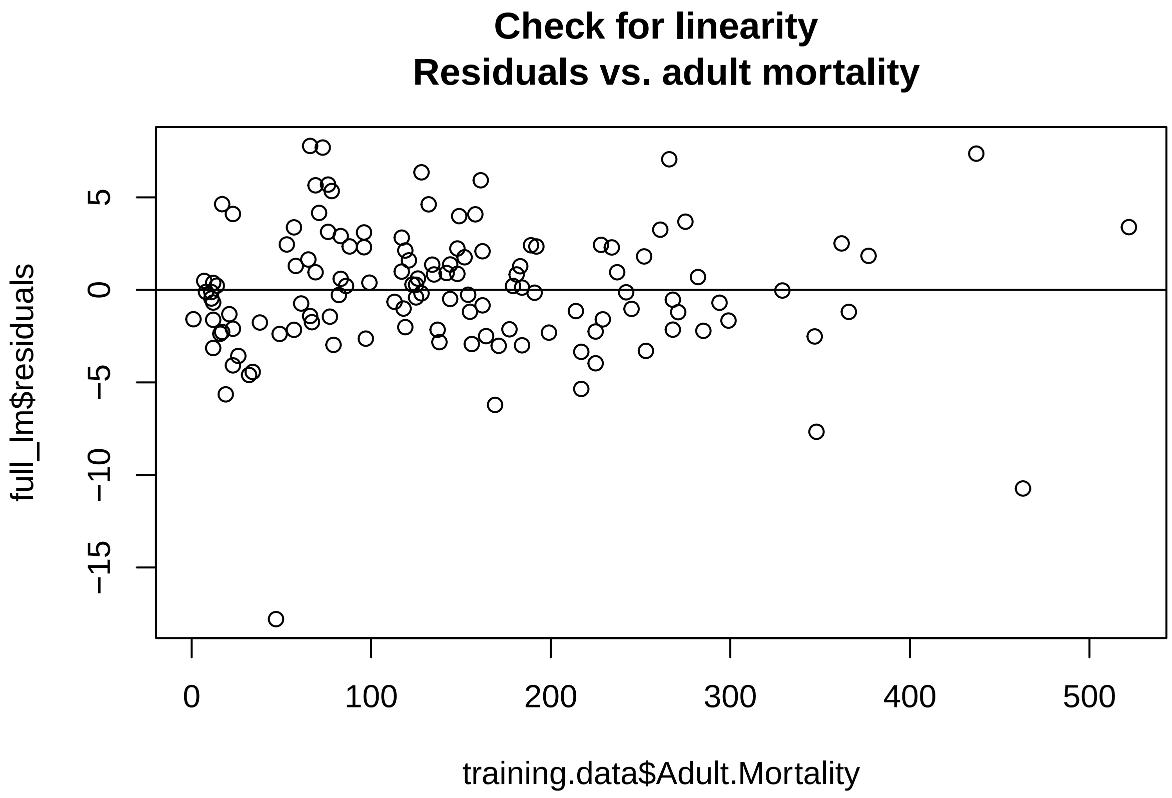
```r
full_lm <- lm(Life.expectancy~., data=training.data)
```

Residual analysis

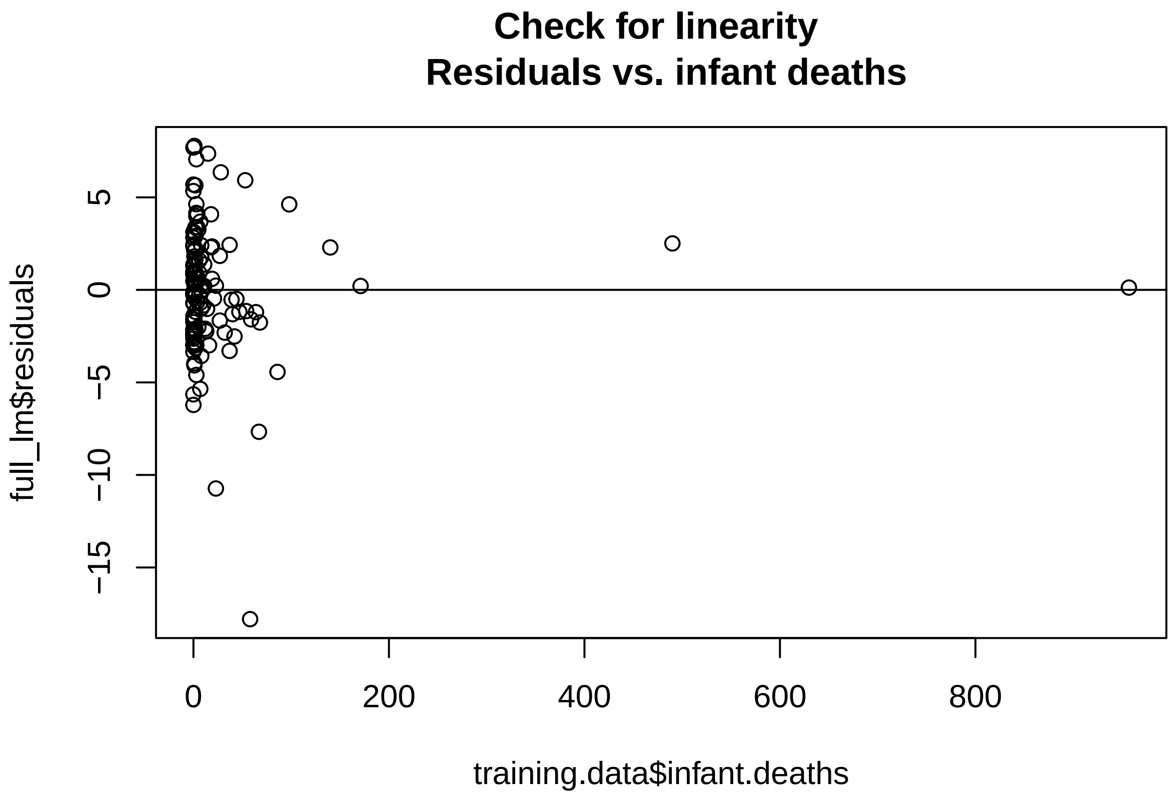
```r
## Check for linearity

plot(x=training.data$Status, y=full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Status")
abline(h=0)
```



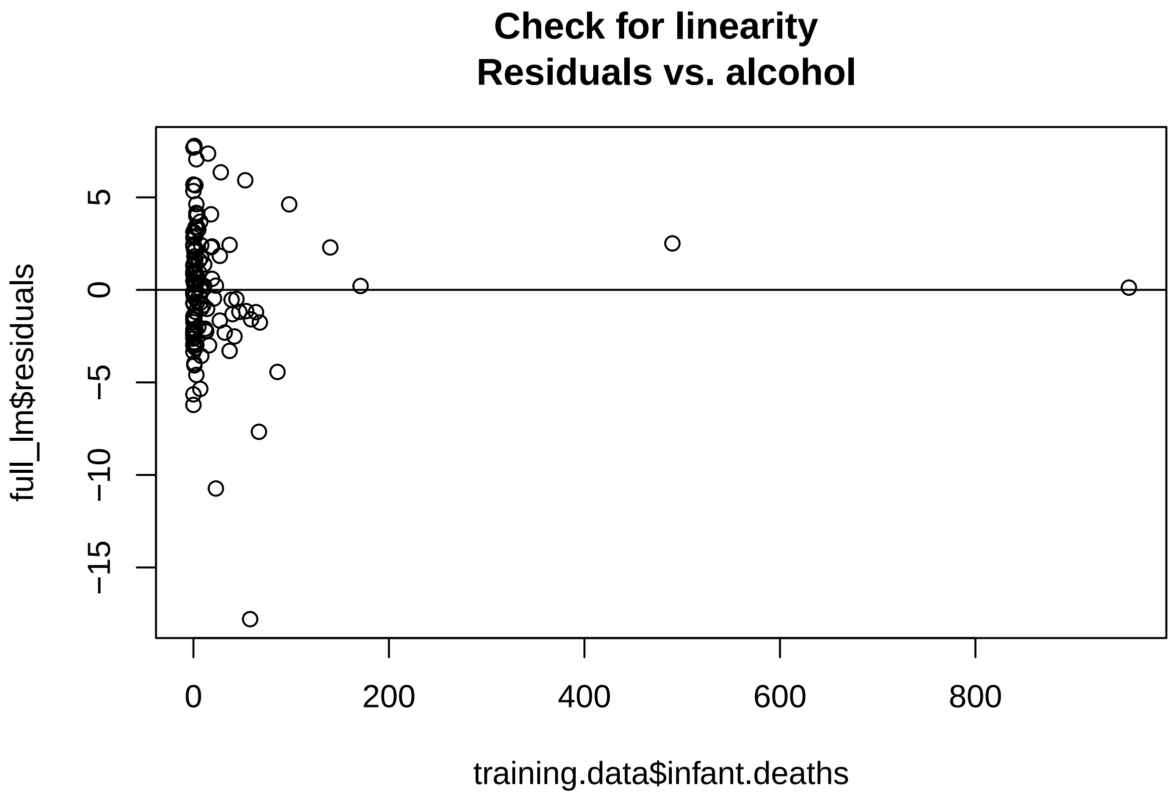
```
plot(training.data$Adult.Mortality, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. adult mortality")
abline(h=0)
```



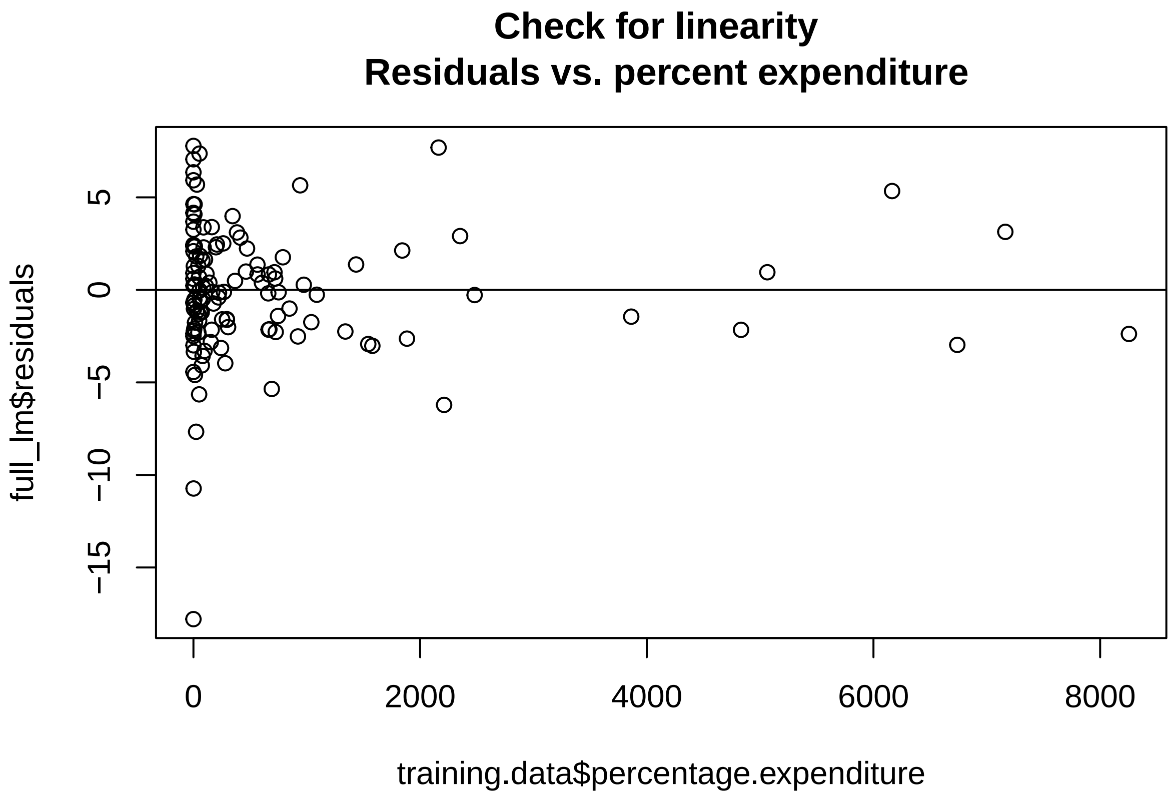
```
plot(training.data$infant.deaths, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. infant deaths")
abline(h=0)
```



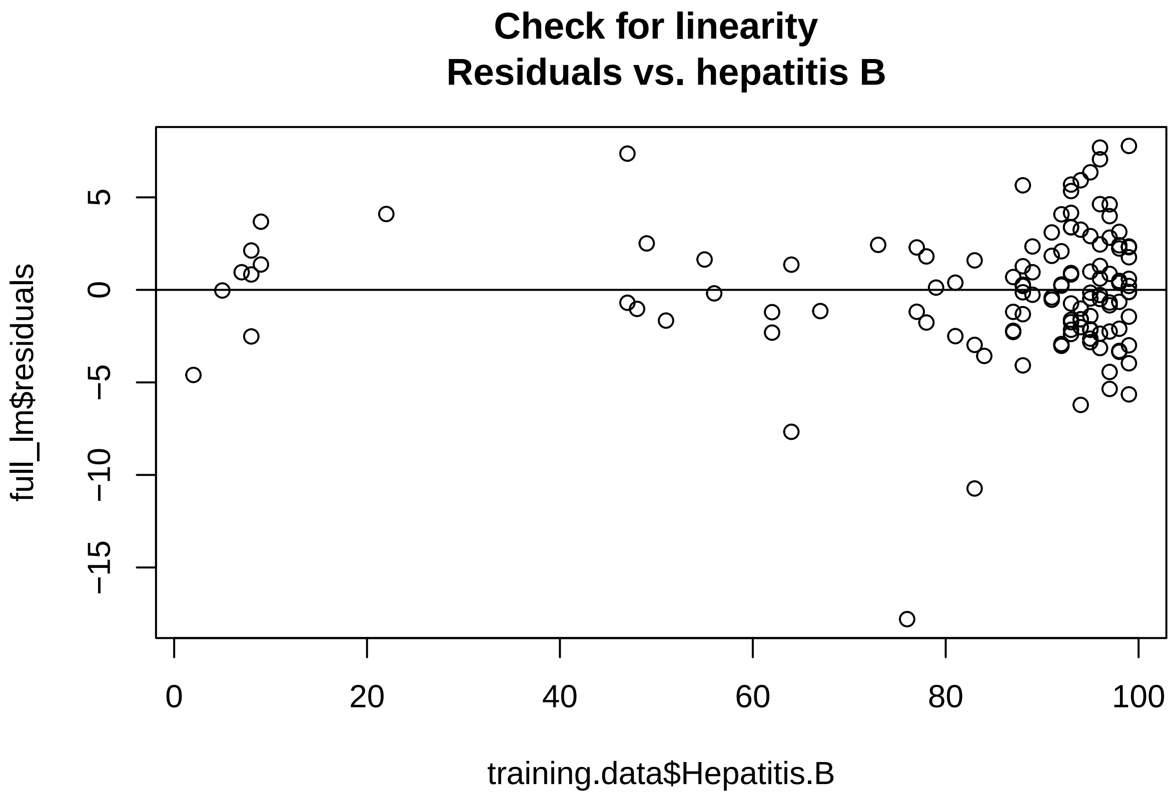
```
plot(training.data$infant.deaths, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. alcohol")
abline(h=0)
```



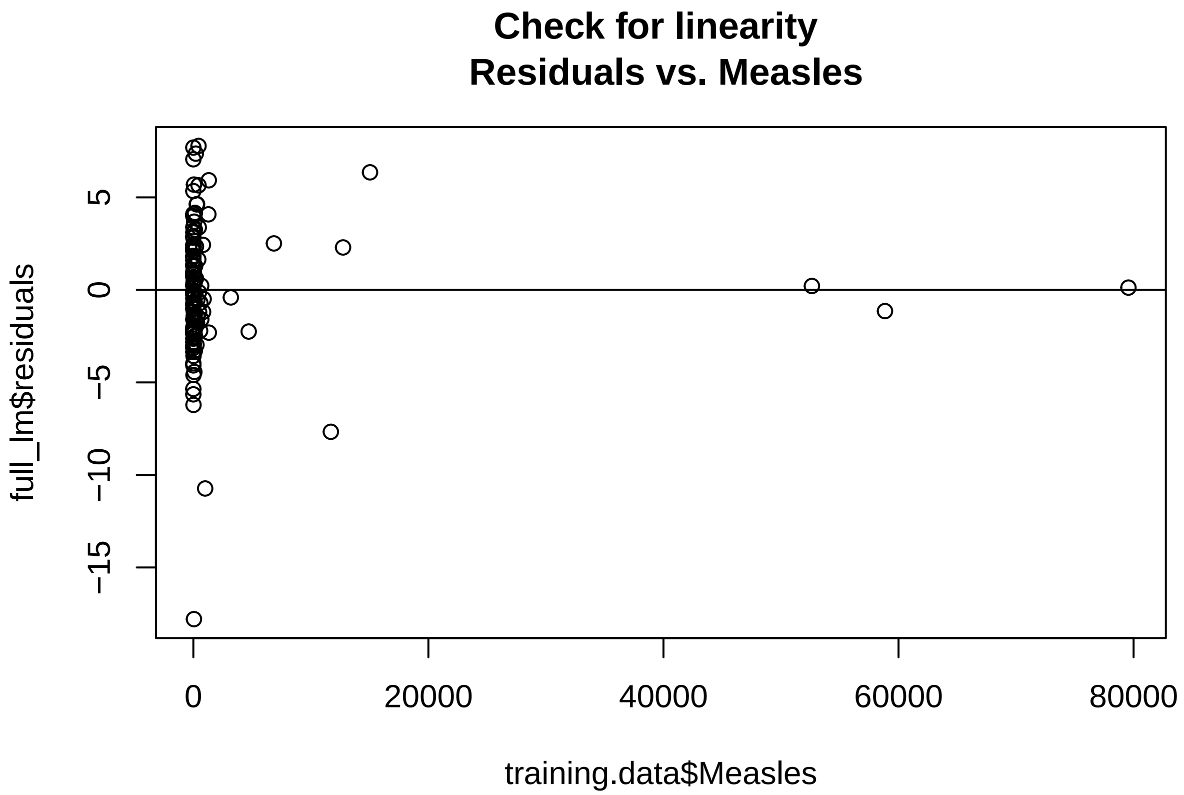
```
plot(training.data$percentage.expenditure, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. percent expenditure")
abline(h=0)
```



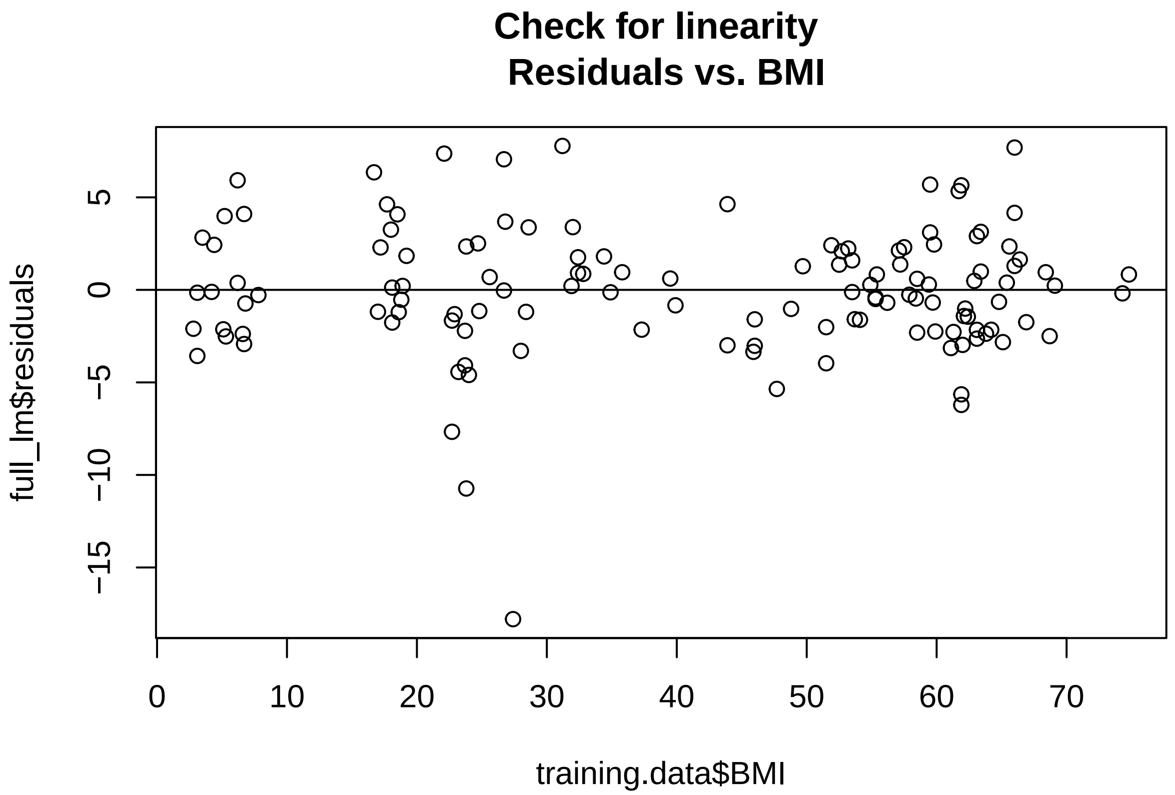
```
plot(training.data$Hepatitis.B, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. hepatitis B")
abline(h=0)
```



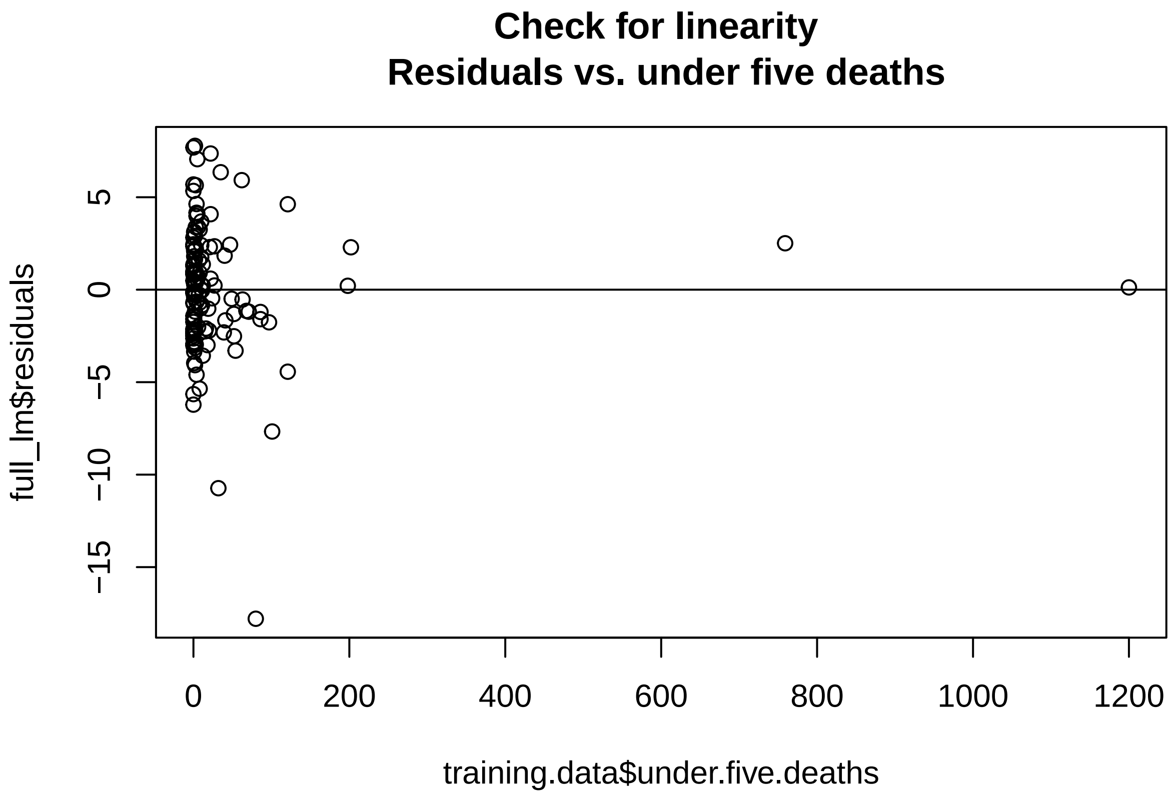
```
plot(training.data$Measles, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Measles")
abline(h=0)
```



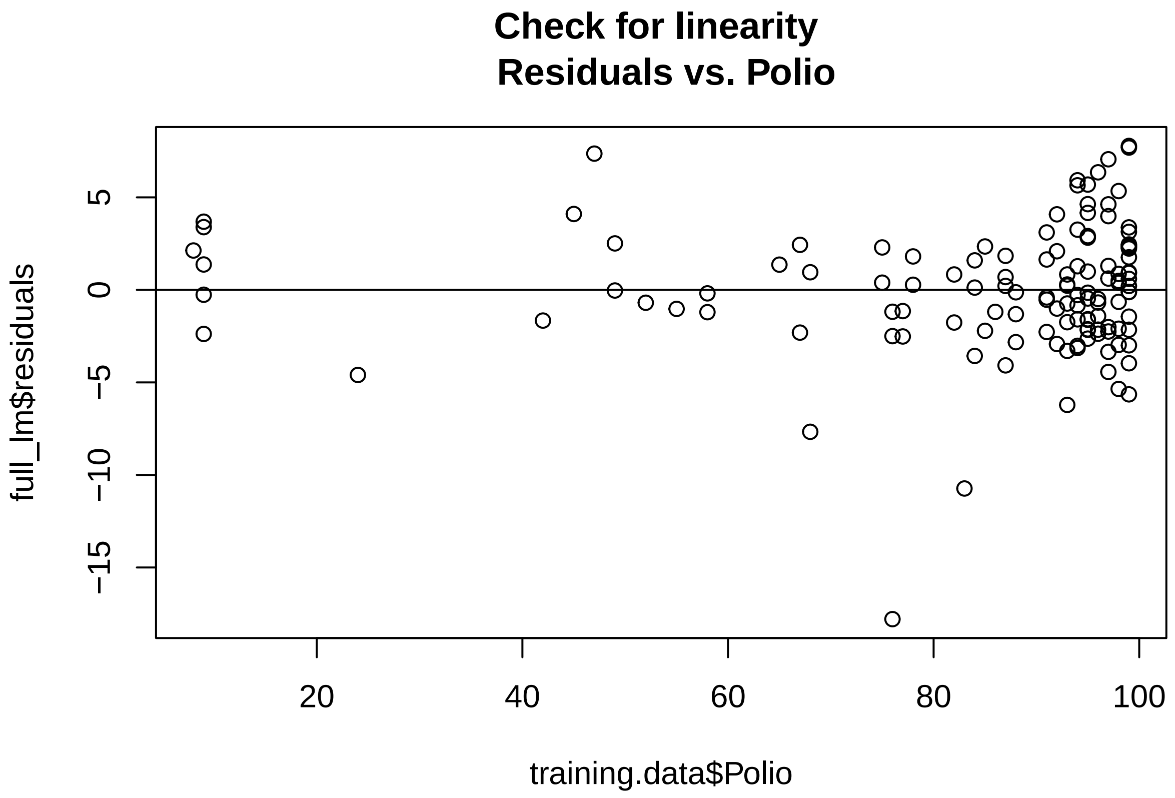
```
plot(training.data$BMI, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. BMI")
abline(h=0)
```



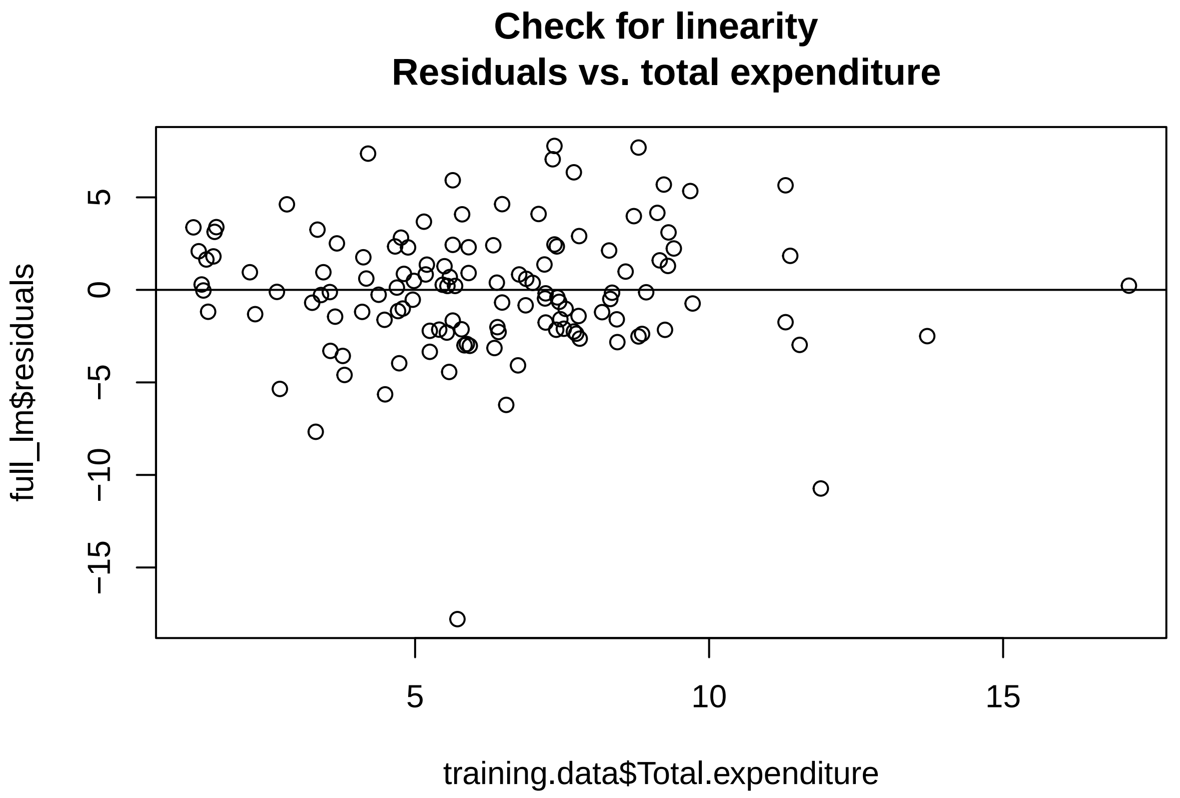
```
plot(training.data$under.five.deaths, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. under five deaths")
abline(h=0)
```



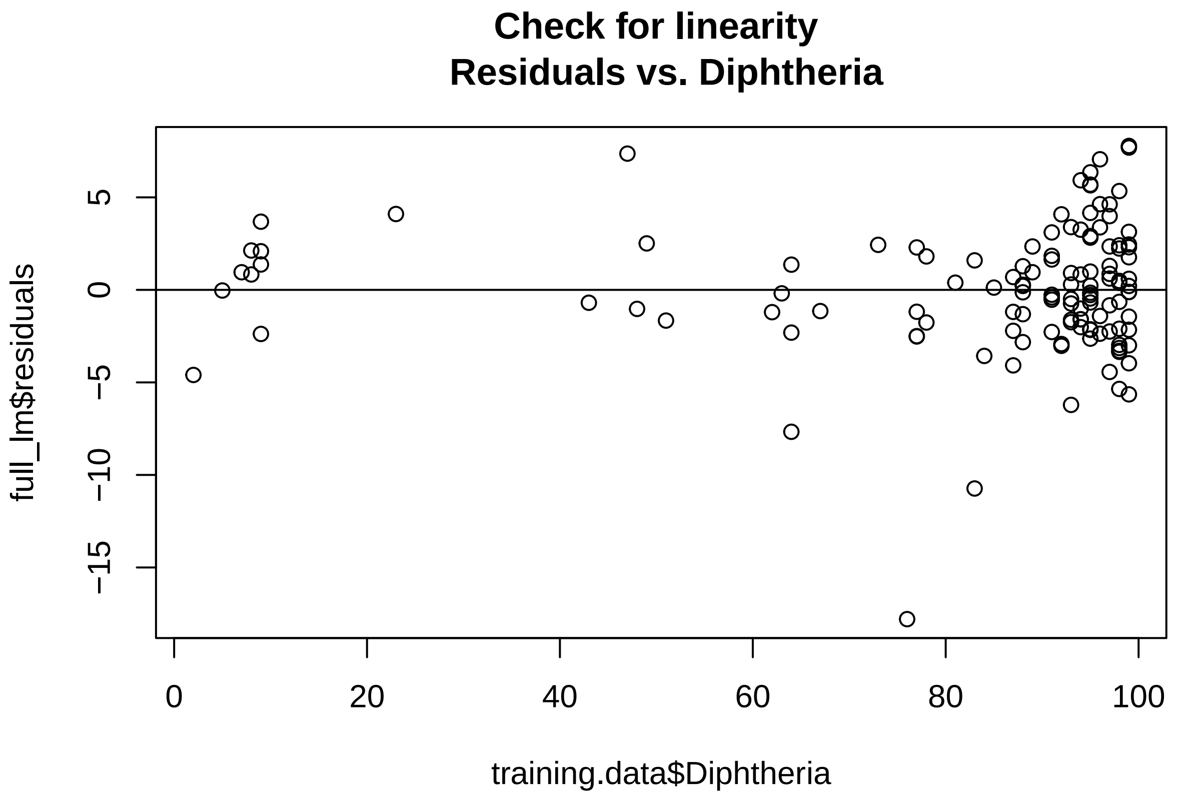
```
plot(training.data$Polio, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Polio")
abline(h=0)
```



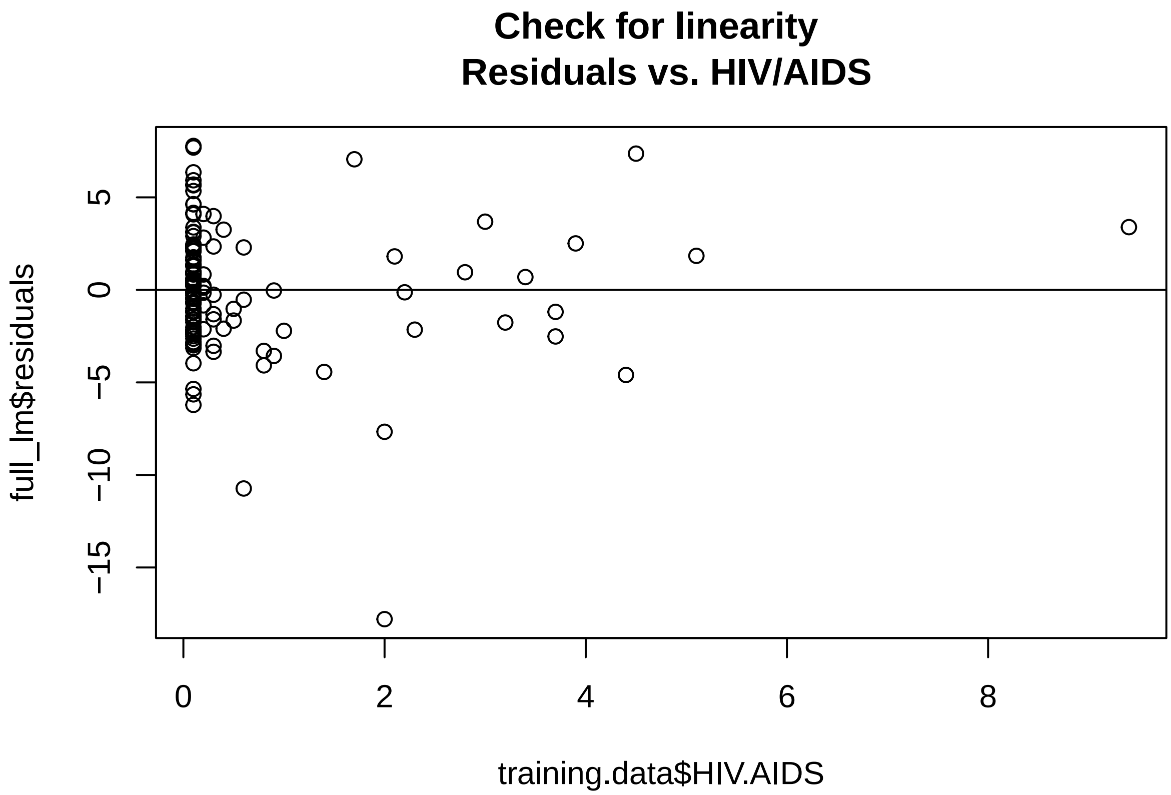
```
plot(training.data$Total.expenditure, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. total expenditure")
abline(h=0)
```



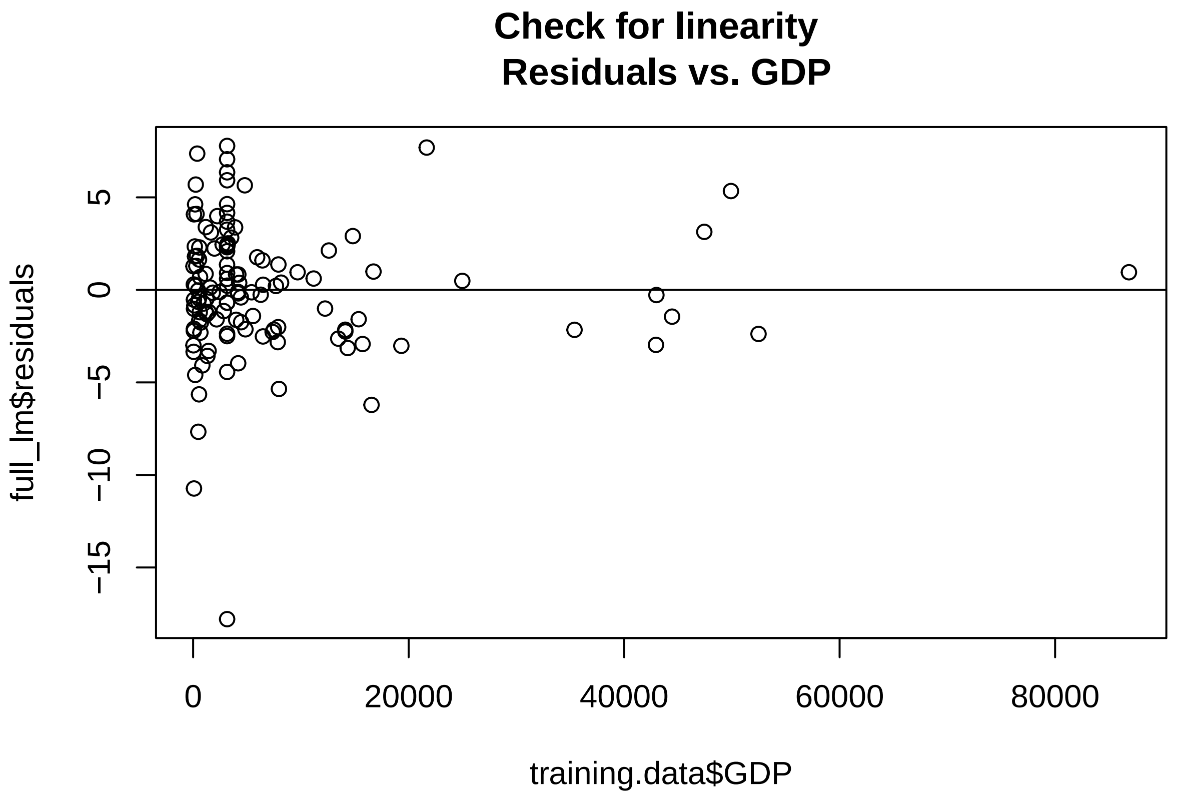
```
plot(training.data$Diphtheria, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Diphtheria")
abline(h=0)
```



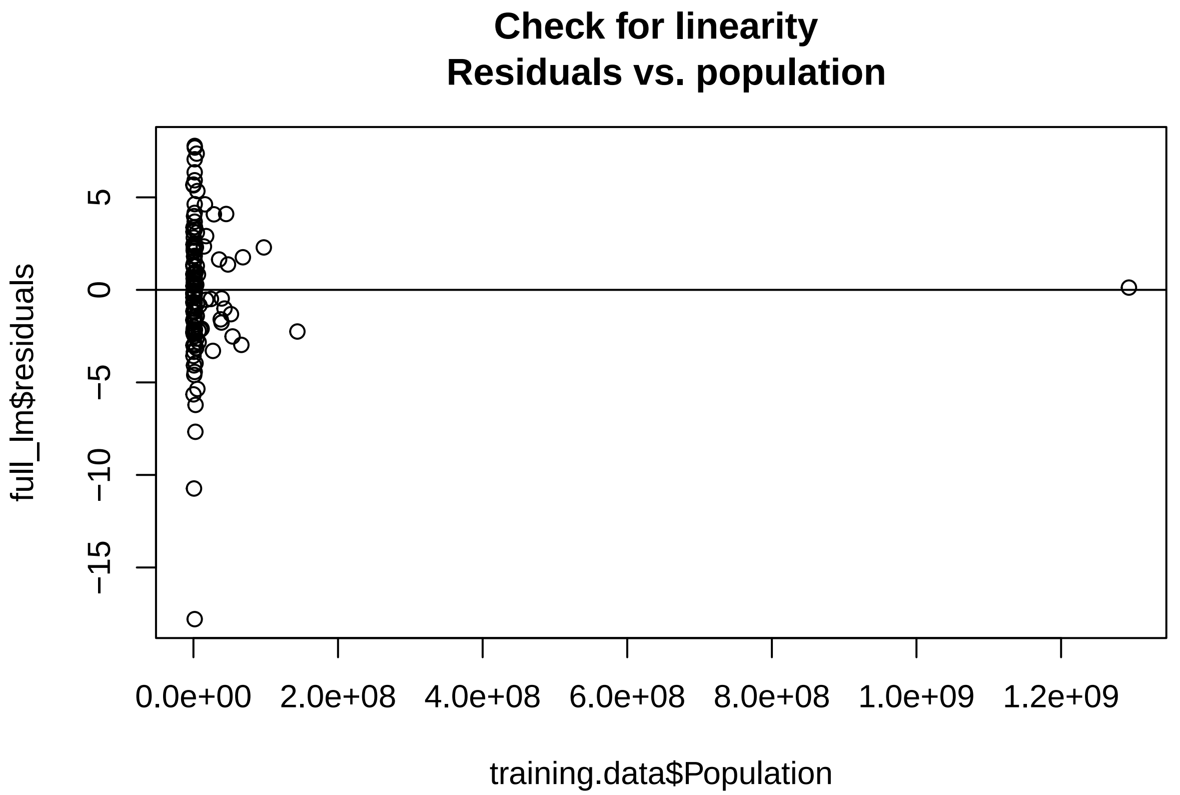
```
plot(training.data$HIV.AIDS, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. HIV/AIDS")
abline(h=0)
```



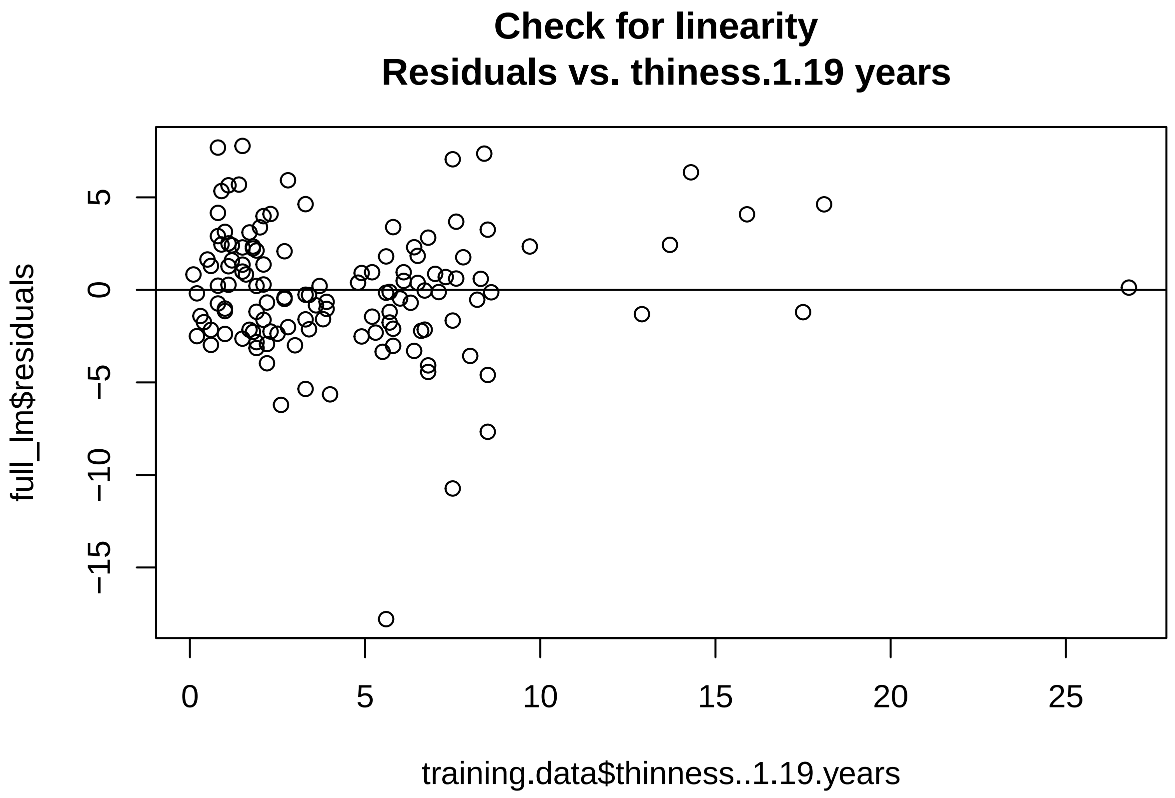
```
plot(training.data$GDP, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. GDP")
abline(h=0)
```



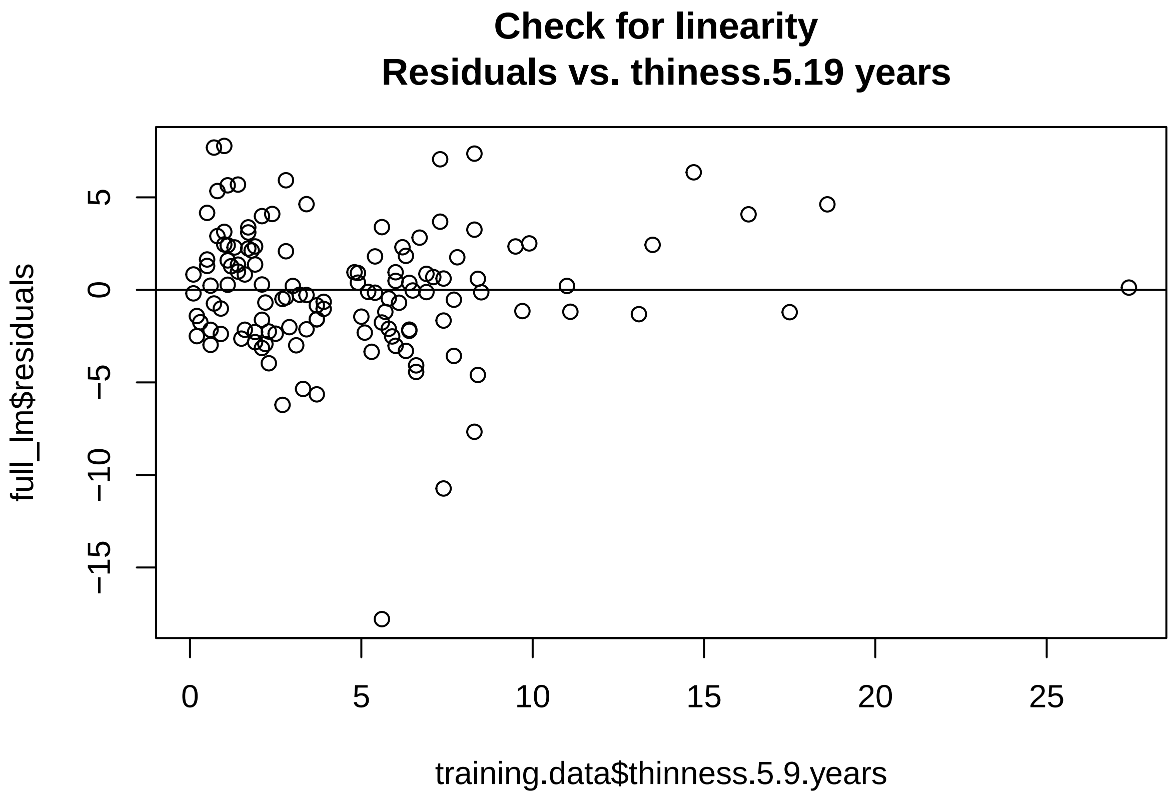
```
plot(training.data$Population, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. population")
abline(h=0)
```



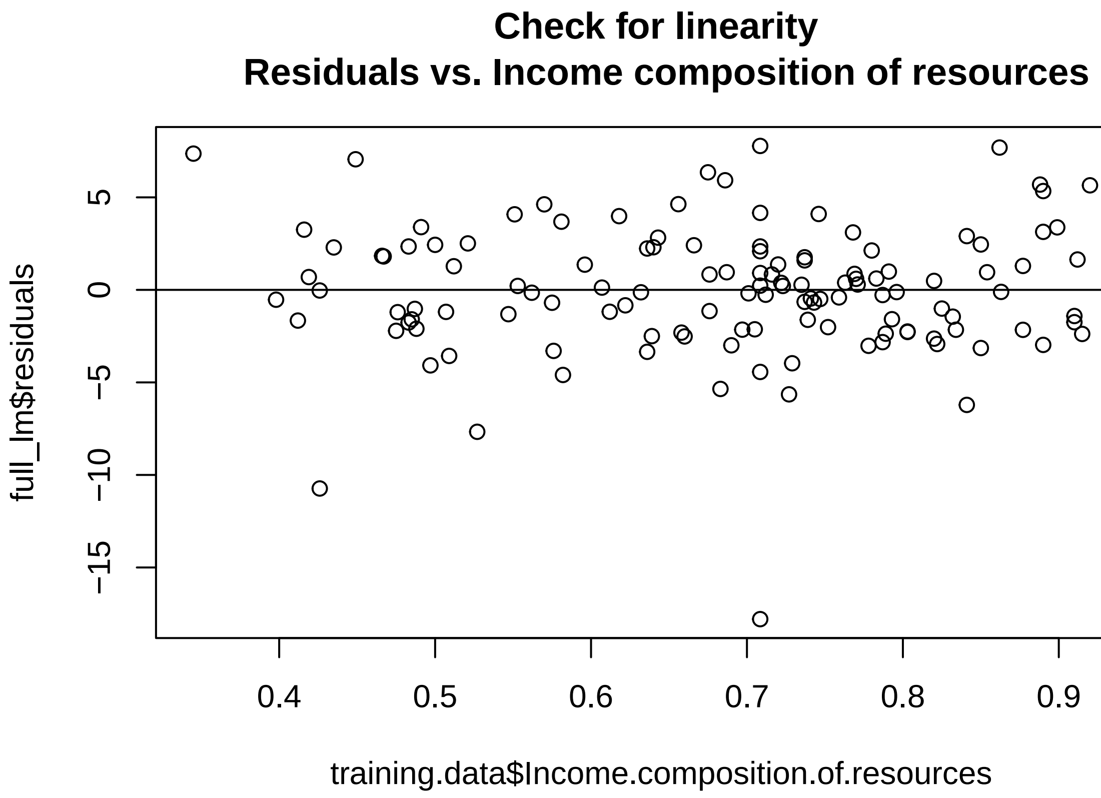
```
plot(training.data$thinness..1.19.years, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. thiness.1.19 years")
abline(h=0)
```



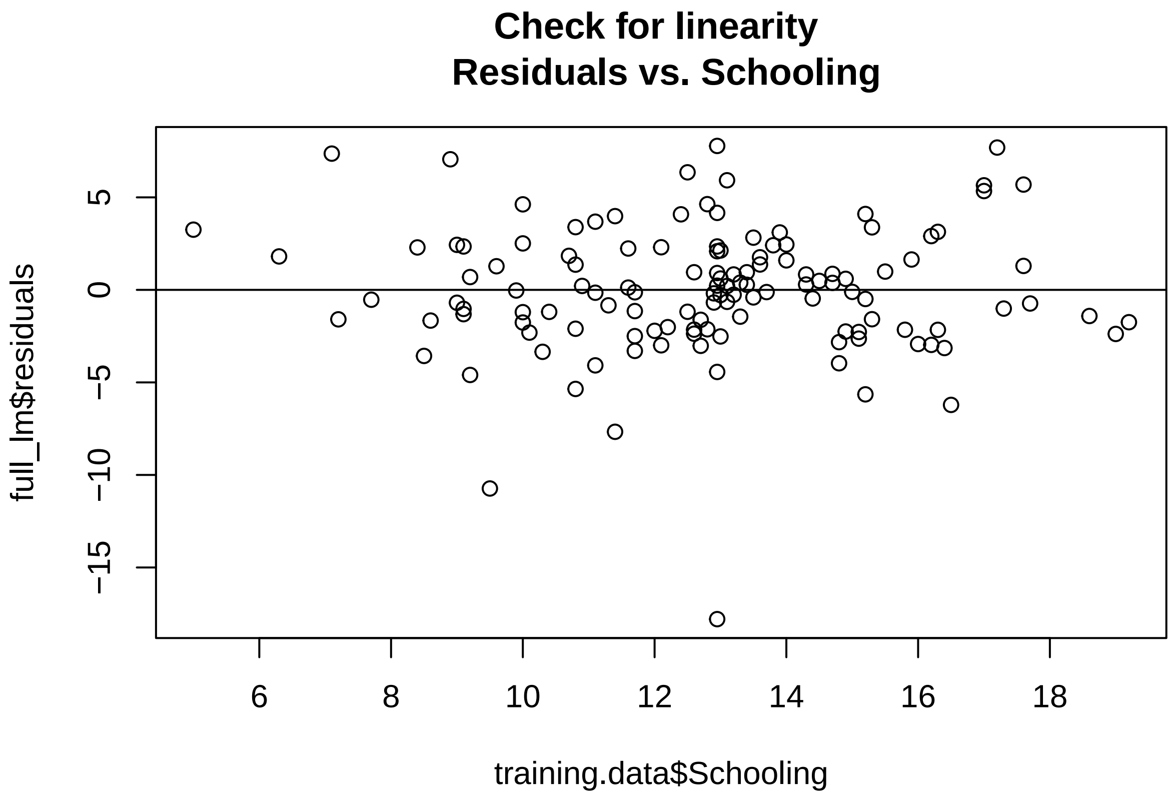
```
plot(training.data$thinness.5.9.years, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. thiness.5.19 years")
abline(h=0)
```



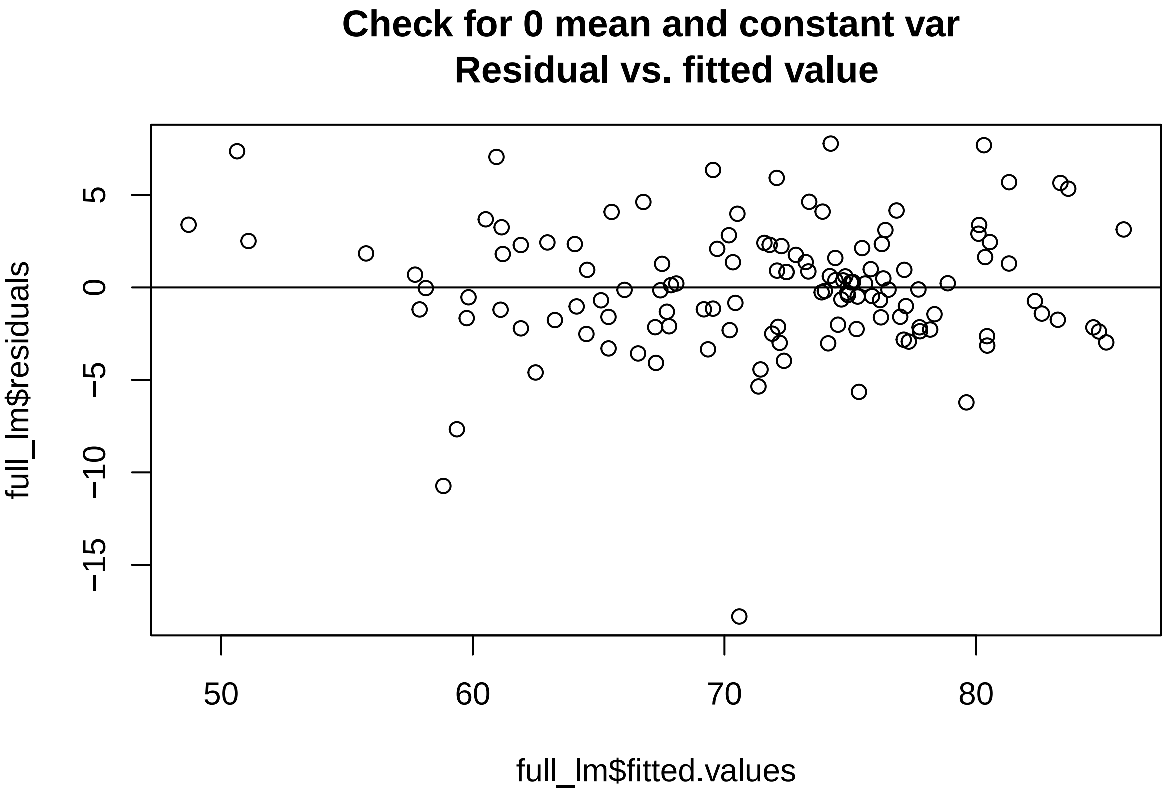
```
plot(training.data$Income.composition.of.resources, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Income composition of resources")
abline(h=0)
```



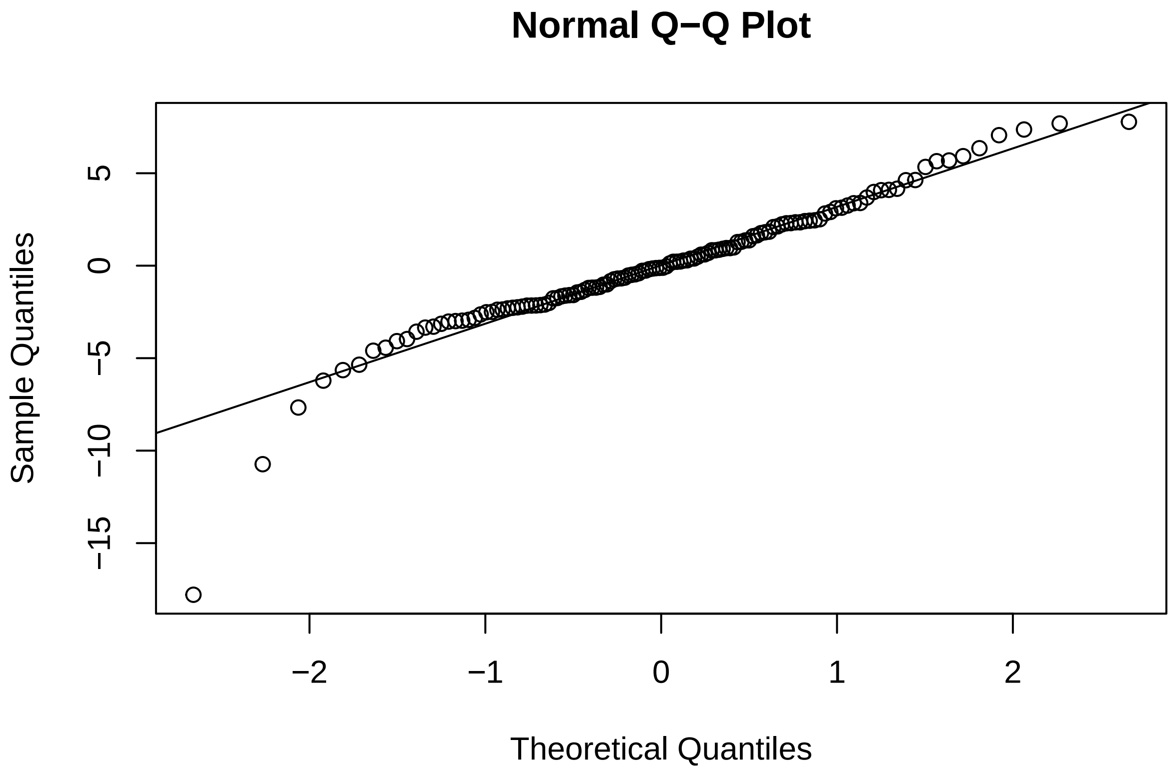
```
plot(training.data$Schooling, full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Schooling")
abline(h=0)
```



```
## Check for zero mean and constant variance of random error
plot(full_lm$fitted.values, full_lm$residuals,
     main = "Check for 0 mean and constant var \n  Residual vs. fitted value")
abline(h=0)
```



```
## Check for normality of random error
qqnorm(full_lm$residuals)
qqline(full_lm$residuals)
```



Definitely needs a transformation as only some of the linearity plots are satisfactory. We are okay with constant variance.

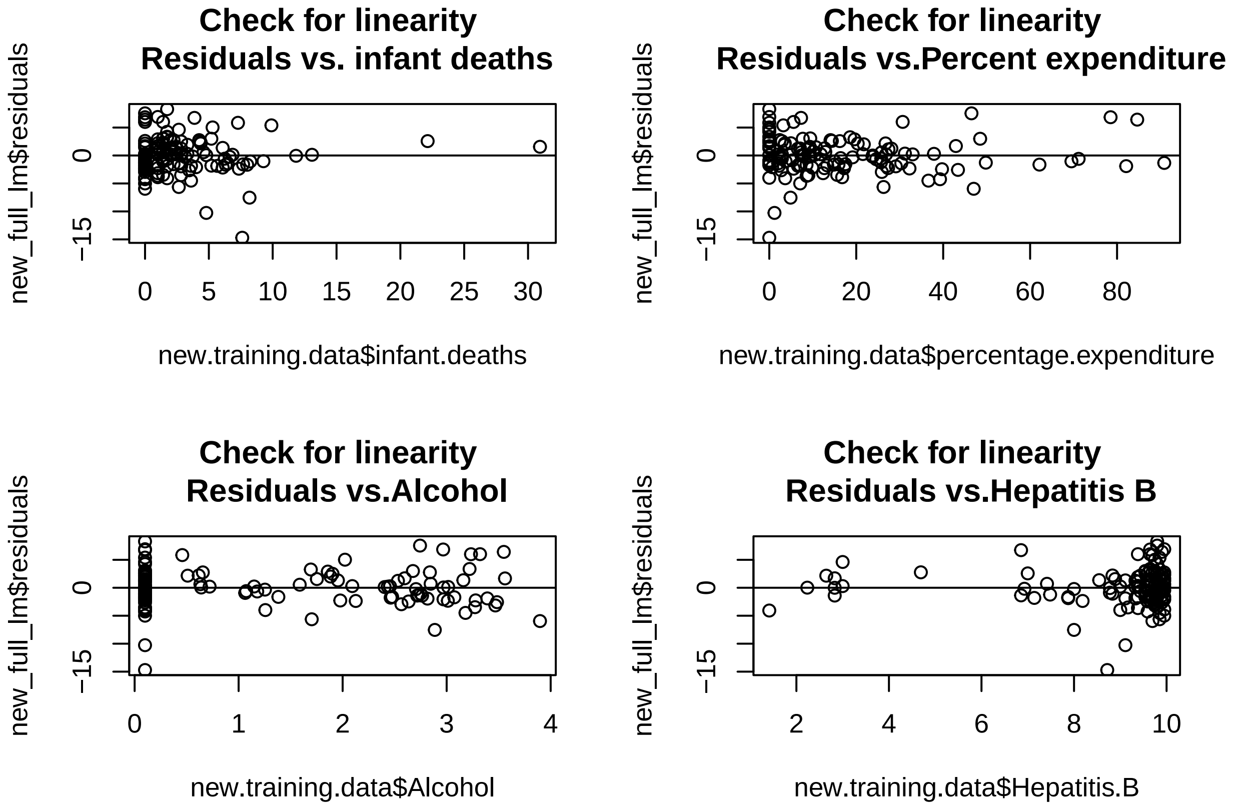
Applying the transformations

```
##### lets log transformation on response variable #####

new.training.data <- training.data
new.training.data$infant.deaths <- new.training.data$infant.deaths^(0.5)

new.training.data$Alcohol <- new.training.data$Alcohol^(0.5)
new.training.data$percentage.expenditure <- new.training.data$percentage.expenditure^(0.5)
new.training.data$Hepatitis.B <- new.training.data$Hepatitis.B^(0.5)

new.training.data$Measles <- new.training.data$Measles^(0.5)

new.training.data$under.five.deaths <- new.training.data$under.five.deaths^(0.5)

new.training.data$Polio <- new.training.data$Polio^(0.5)

new.training.data$Diphtheria<- new.training.data$Diphtheria^(0.5)

new.training.data$HIV.AIDS <- new.training.data$HIV.AIDS^(0.5)
```

```r
new.training.data$GDP <- new.training.data$GDP^(0.5)

new.training.data$Population <- new.training.data$Population^(0.5)


new.training.data$Measles <- new.training.data$Measles^(0.5)

new.training.data$thinness..1.19.years <- new.training.data$thinness..1.19.years^(0.5)

new.training.data$thinness.5.9.years <- new.training.data$thinness.5.9.years^(0.5)

new_full_lm <- lm(Life.expectancy ~. , data = new.training.data)
summary(new_full_lm )
```

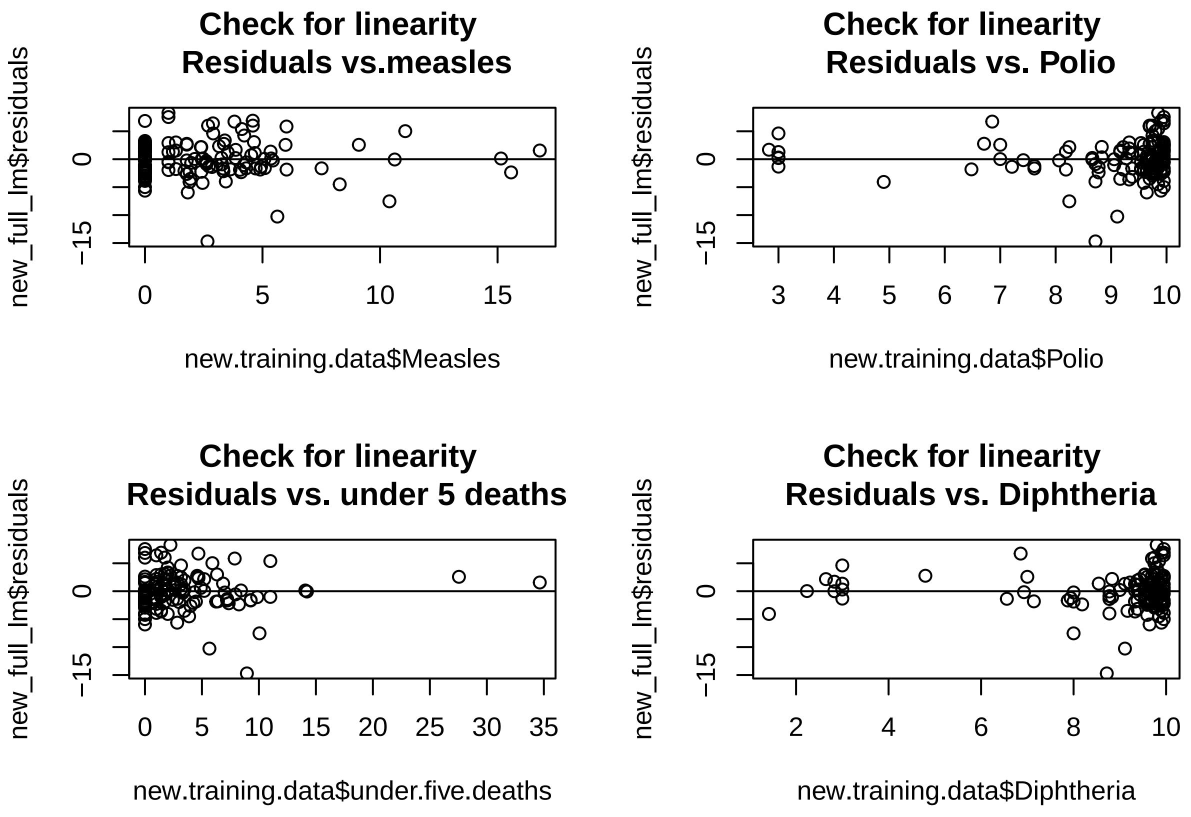
```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = new.training.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7063  -1.8244  -0.0269   1.9422   8.2850
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     64.3986943  4.9449310  13.023  < 2e-16 ***
## Statusyes                       -1.6957714  1.1237354  -1.509 0.134206
## Adult.Mortality                 -0.0173878  0.0046392  -3.748 0.000288 ***
## infant.deaths                    1.4079174  1.3622143   1.034 0.303655
## Alcohol                          0.2239293  0.3393691   0.660 0.510761
## percentage.expenditure           0.0449840  0.0372694   1.207 0.230071
## Hepatitis.B                     -0.0484411  0.3703135  -0.131 0.896168
## Measles                          0.1774388  0.1521522   1.166 0.246105
## BMI                             -0.0068524  0.0205970  -0.333 0.740012
## under.five.deaths               -1.4182176  1.1554574  -1.227 0.222338
## Polio                           -0.1175258  0.3068245  -0.383 0.702444
## Total.expenditure                0.0591599  0.1448075   0.409 0.683685
## Diphtheria                       0.3366212  0.3806582   0.884 0.378491
## HIV.AIDS                        -3.7587663  0.8945574  -4.202 5.47e-05 ***
## GDP                             -0.0028250  0.0136520  -0.207 0.836455
## Population                       0.0001081  0.0001338   0.808 0.420705
## thinness..1.19.years            -0.1676508  1.0140479  -0.165 0.868995
## thinness.5.9.years              -1.1891533  1.0657718  -1.116 0.266998
## Income.composition.of.resources 20.4447935  7.7396298   2.642 0.009477 **
## Schooling                       -0.0380177  0.3247116  -0.117 0.907013
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.608 on 108 degrees of freedom
## Multiple R-squared:  0.8423, Adjusted R-squared:  0.8146
## F-statistic: 30.37 on 19 and 108 DF,  p-value: < 2.2e-16
```

```r
## Check for linearity
par(mfcol = c(2, 2))
plot(new.training.data$infant.deaths, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. infant deaths")
abline(h=0)


plot(new.training.data$Alcohol, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs.Alcohol")
abline(h=0)


plot(new.training.data$percentage.expenditure, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs.Percent expenditure")
abline(h=0)


plot(new.training.data$Hepatitis.B, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs.Hepatitis B")
abline(h=0)
```



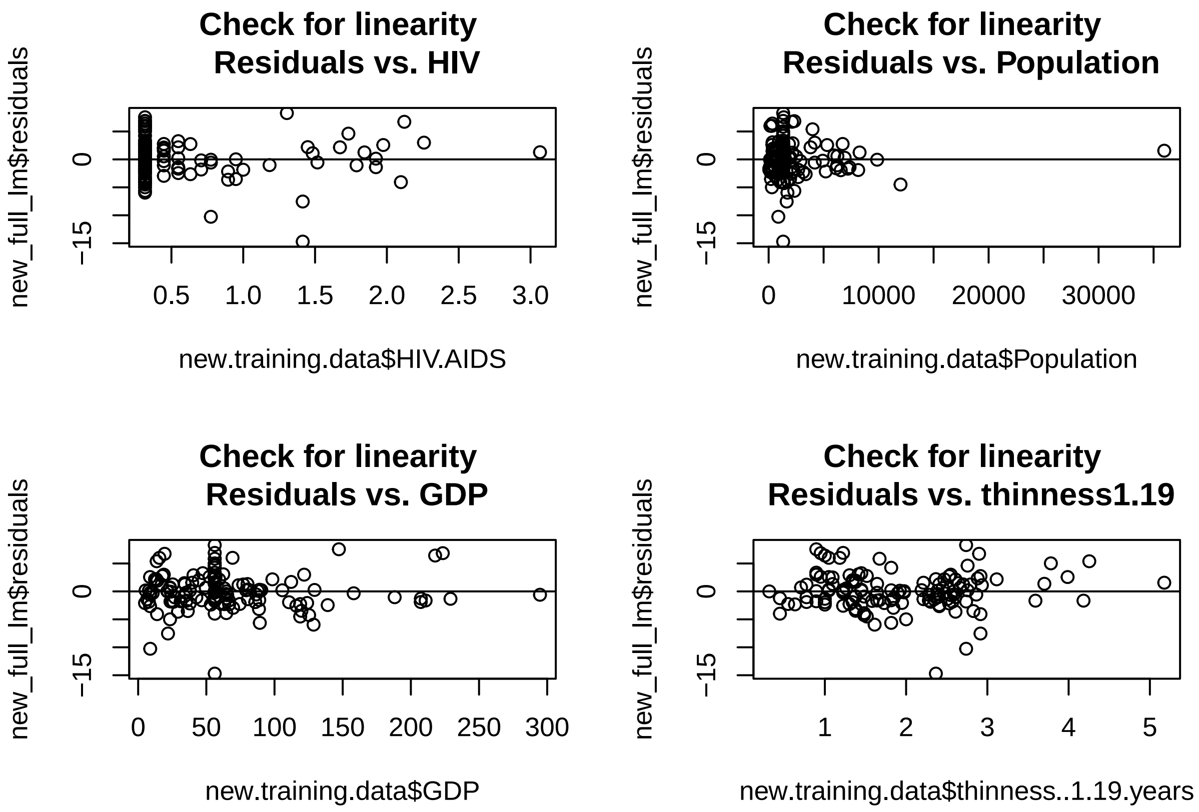
```r
plot(new.training.data$Measles, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs.measles")
abline(h=0)
```

```
plot(new.training.data$under.five.deaths, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. under 5 deaths")
abline(h=0)


plot(new.training.data$Polio, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Polio")
abline(h=0)


plot(new.training.data$Diphtheria, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Diphtheria")
abline(h=0)
```



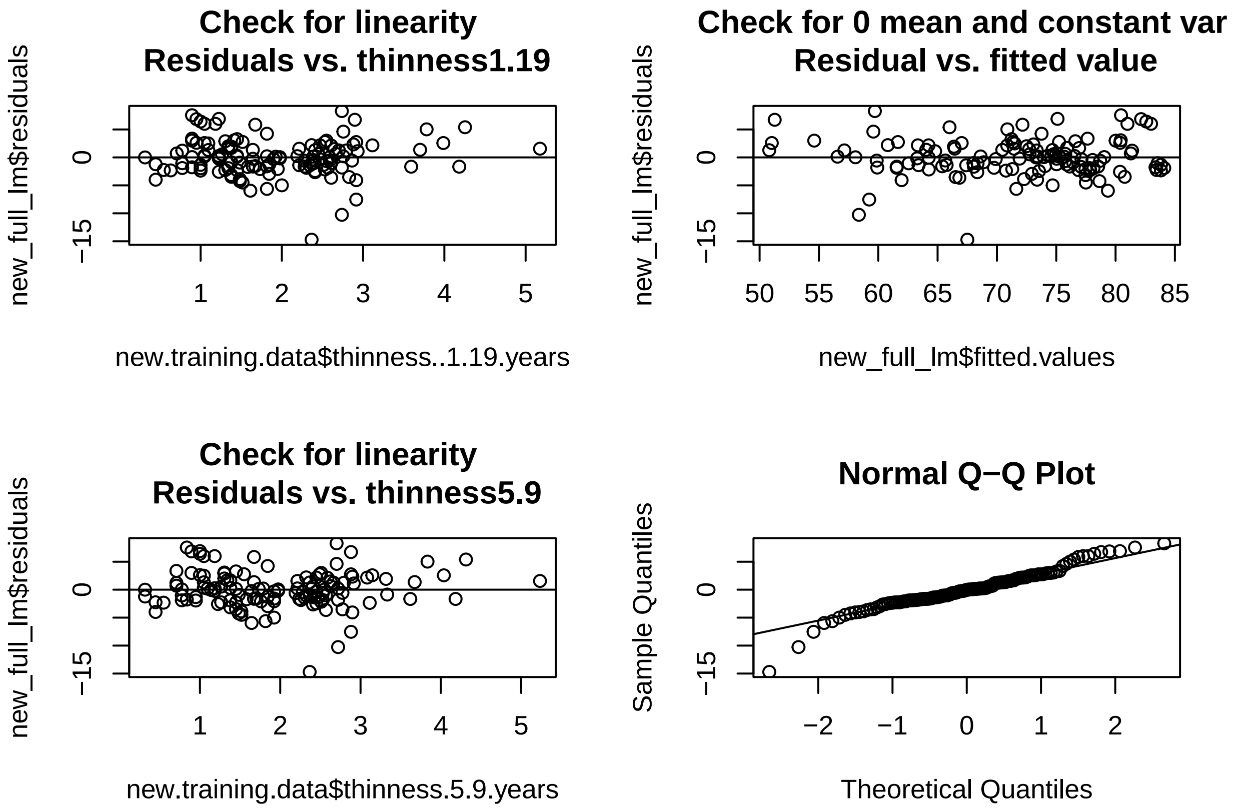
```
plot(new.training.data$HIV.AIDS, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. HIV")
abline(h=0)

plot(new.training.data$GDP, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. GDP")
abline(h=0)

plot(new.training.data$Population, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. Population")
abline(h=0)
```

```
plot(new.training.data$thinness..1.19.years, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. thinness1.19")
abline(h=0)
```



```
plot(new.training.data$thinness..1.19.years, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. thinness1.19")
abline(h=0)

plot(new.training.data$thinness.5.9.years, new_full_lm$residuals,
     main = "Check for linearity \n Residuals vs. thinness5.9")
abline(h=0)
















# Check for zero mean and constant variance of random error
plot(new_full_lm$fitted.values, new_full_lm$residuals,
     main = "Check for 0 mean and constant var \n  Residual vs. fitted value")
```

```r
abline(h=0)

## Check for normality of random error
qqnorm(new_full_lm$residuals)
qqline(new_full_lm$residuals)
```



So, the new training model is

```r
summary(new_full_lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = new.training.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7063  -1.8244  -0.0269   1.9422   8.2850
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     64.3986943  4.9449310  13.023  < 2e-16 ***
## Statusyes                       -1.6957714  1.1237354  -1.509 0.134206
## Adult.Mortality                 -0.0173878  0.0046392  -3.748 0.000288 ***
## infant.deaths                    1.4079174  1.3622143   1.034 0.303655
## Alcohol                          0.2239293  0.3393691   0.660 0.510761
## percentage.expenditure           0.0449840  0.0372694   1.207 0.230071
## Hepatitis.B                     -0.0484411  0.3703135  -0.131 0.896168
## Measles                          0.1774388  0.1521522   1.166 0.246105
## BMI                             -0.0068524  0.0205970  -0.333 0.740012
## under.five.deaths               -1.4182176  1.1554574  -1.227 0.222338
## Polio                           -0.1175258  0.3068245  -0.383 0.702444
## Total.expenditure                0.0591599  0.1448075   0.409 0.683685
## Diphtheria                       0.3366212  0.3806582   0.884 0.378491
## HIV.AIDS                        -3.7587663  0.8945574  -4.202 5.47e-05 ***
## GDP                             -0.0028250  0.0136520  -0.207 0.836455
## Population                       0.0001081  0.0001338   0.808 0.420705
## thinness..1.19.years            -0.1676508  1.0140479  -0.165 0.868995
## thinness.5.9.years              -1.1891533  1.0657718  -1.116 0.266998
## Income.composition.of.resources 20.4447935  7.7396298   2.642 0.009477 **
## Schooling                       -0.0380177  0.3247116  -0.117 0.907013
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.608 on 108 degrees of freedom
## Multiple R-squared:  0.8423, Adjusted R-squared:  0.8146
## F-statistic: 30.37 on 19 and 108 DF,  p-value: < 2.2e-16
```

Adjusted R _squared is pretty high.Residual standard error is also low.

We will work on this model.