Regression Analysis on Factors Influencing Student Exam Scores

Yingqi Jiang

2024-11-04

```
df <- read.csv("StudentPerformanceFactors.csv", header=TRUE, sep=",")
summary(df)
```

```
##  Hours_Studied     Attendance     Parental_Involvement Access_to_Resources
##  Min.   : 1.00   Min.   : 60.00   Length:6607          Length:6607
##  1st Qu.:16.00   1st Qu.: 70.00   Class :character     Class :character
##  Median :20.00   Median : 80.00   Mode  :character     Mode  :character
##  Mean   :19.98   Mean   : 79.98
##  3rd Qu.:24.00   3rd Qu.: 90.00
##  Max.   :44.00   Max.   :100.00
##  Extracurricular_Activities  Sleep_Hours     Previous_Scores
##  Length:6607                Min.   : 4.000   Min.   : 50.00
##  Class :character           1st Qu.: 6.000   1st Qu.: 63.00
##  Mode  :character           Median : 7.000   Median : 75.00
##                             Mean   : 7.029   Mean   : 75.07
##                             3rd Qu.: 8.000   3rd Qu.: 88.00
##                             Max.   :10.000   Max.   :100.00
##  Motivation_Level   Internet_Access    Tutoring_Sessions Family_Income
##  Length:6607        Length:6607        Min.   :0.000     Length:6607
##  Class :character   Class :character   1st Qu.:1.000     Class :character
##  Mode  :character   Mode  :character   Median :1.000     Mode  :character
##                                        Mean   :1.494
##                                        3rd Qu.:2.000
##                                        Max.   :8.000
##  Teacher_Quality    School_Type        Peer_Influence     Physical_Activity
##  Length:6607        Length:6607        Length:6607        Min.   :0.000
##  Class :character   Class :character   Class :character   1st Qu.:2.000
##  Mode  :character   Mode  :character   Mode  :character   Median :3.000
##                                                           Mean   :2.968
##                                                           3rd Qu.:4.000
##                                                           Max.   :6.000
##  Learning_Disabilities Parental_Education_Level Distance_from_Home
##  Length:6607           Length:6607              Length:6607
##  Class :character      Class :character         Class :character
##  Mode  :character      Mode  :character         Mode  :character
##
##
##
##     Gender            Exam_Score
##  Length:6607        Min.   : 55.00
##  Class :character   1st Qu.: 65.00
##  Mode  :character   Median : 67.00
##                     Mean   : 67.24
```

```
##                      3rd Qu.: 69.00
##                      Max.   :101.00
```

```r
# Convert 'parent_education_level' to a factor
df$Peer_Influence <- factor(df$Peer_Influence,
                                      levels = c("Positive", "Neutral", "Negative"))

# Fit the linear regression model
model1 <- lm(Exam_Score ~ Hours_Studied + Attendance + Peer_Influence, data = df)

summary(model1)
```

```
##
## Call:
## lm(formula = Exam_Score ~ Hours_Studied + Attendance + Peer_Influence,
##     data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.465 -1.289 -0.171  0.984 32.146
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             46.014370   0.252222 182.436  < 2e-16 ***
## Hours_Studied            0.292431   0.005354  54.614  < 2e-16 ***
## Attendance               0.197527   0.002779  71.086  < 2e-16 ***
## Peer_InfluenceNeutral   -0.516871   0.072124  -7.166 8.53e-13 ***
## Peer_InfluenceNegative -1.031788   0.086668 -11.905  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.607 on 6602 degrees of freedom
## Multiple R-squared:  0.5513, Adjusted R-squared:  0.551
## F-statistic:  2028 on 4 and 6602 DF,  p-value: < 2.2e-16
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
```

```
##     as.Date, as.Date.numeric
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

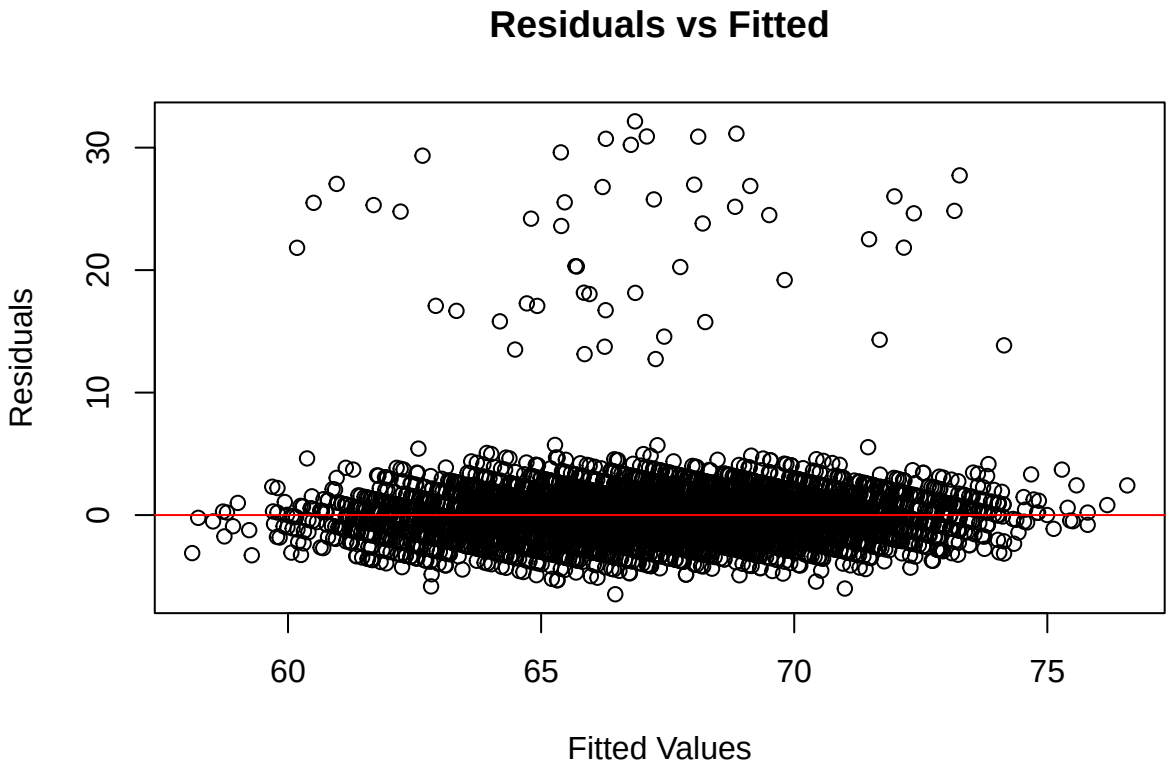
```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
# DW test for autocorrelation
dwtest(model1)
```

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 2.0033, p-value = 0.5527
## alternative hypothesis: true autocorrelation is greater than 0
```

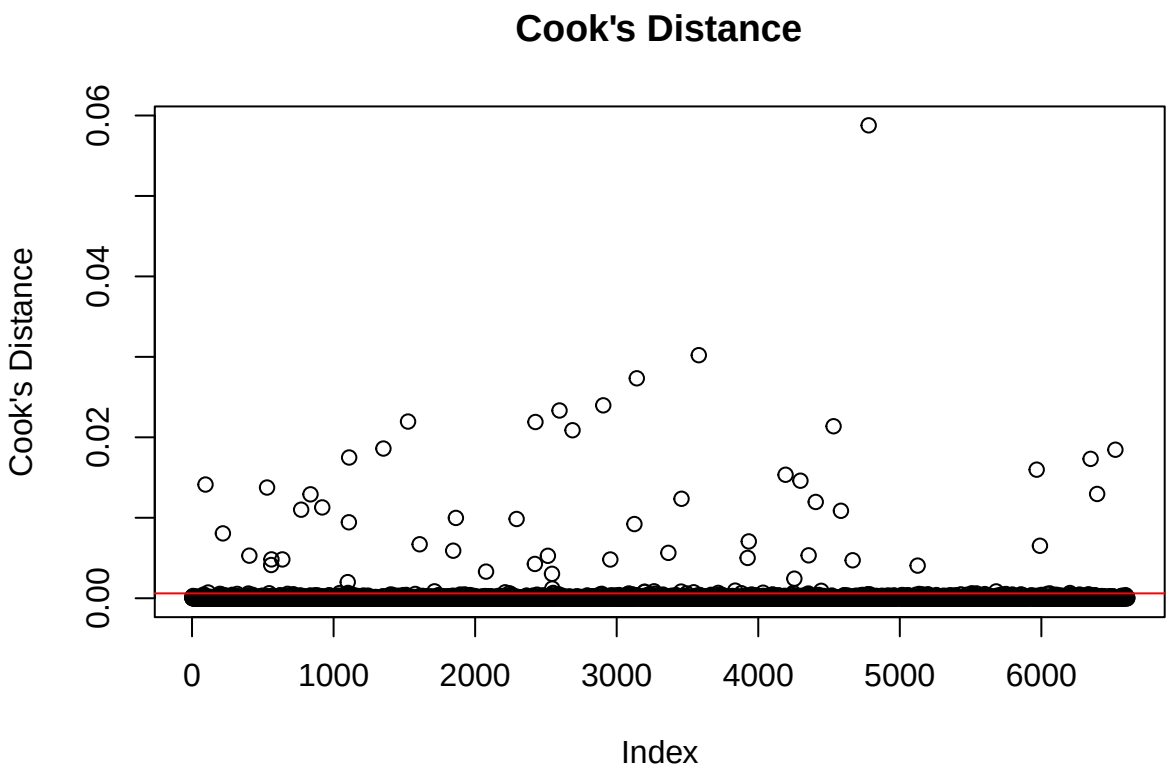
```
# check heteroscedasticity
plot(model1$residuals ~ model1$fitted.values, main="Residuals vs Fitted", xlab = "Fitted Values", ylab =
abline(h=0, col="red")
```




```
# VIF for multicollinearity
vif(model1)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## Hours_Studied  1.000232  1        1.000116
## Attendance     1.000888  1        1.000444
## Peer_Influence 1.000928  2        1.000232
```

```
# Influential points using Cook's distance
cooksd <- cooks.distance(model1)
plot(cooksd, main="Cook's Distance", ylab="Cook's Distance")
abline(h=4/length(cooksd), col="red")
```

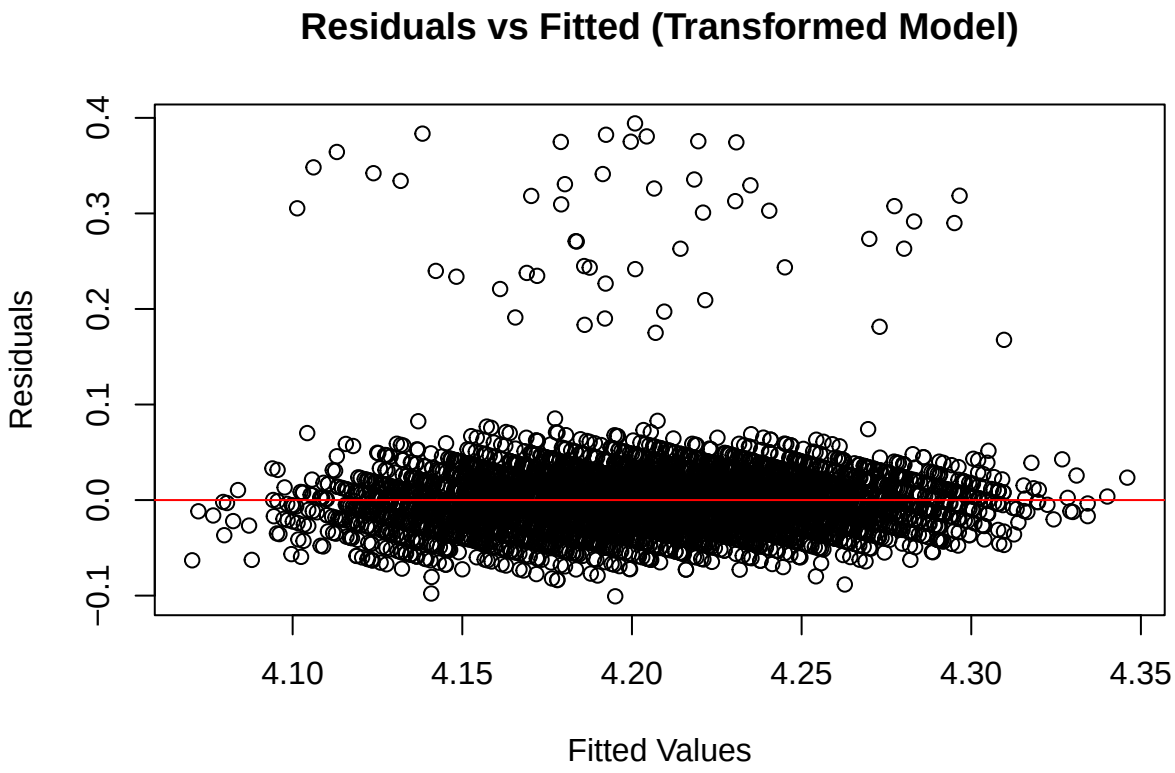



```r
# Apply a log transformation to the response variable
df$ExamScore_log <- log(df$Exam_Score)

# Refit the model using the transformed response variable
model_transformed <- lm(ExamScore_log ~ Hours_Studied + Attendance + Peer_Influence, data = df)

# Summarize the transformed model
summary(model_transformed)
```

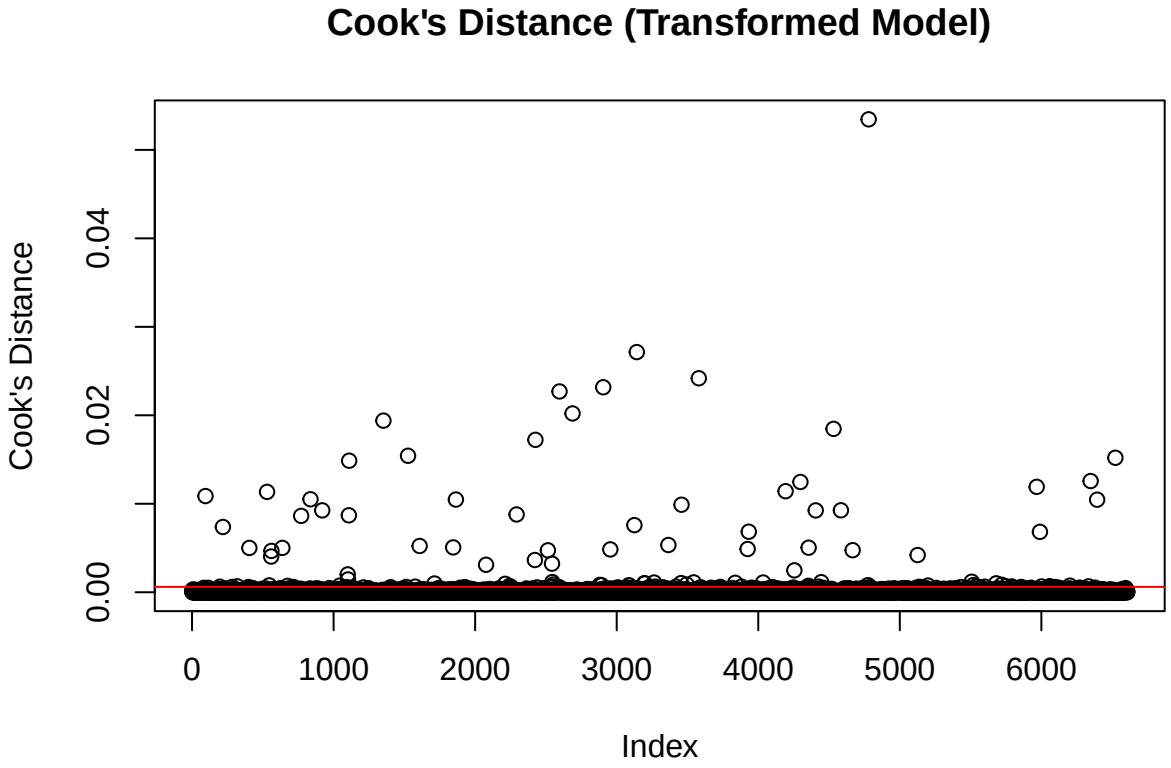
```
##
## Call:
## lm(formula = ExamScore_log ~ Hours_Studied + Attendance + Peer_Influence,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10073 -0.01866 -0.00186  0.01517  0.39422
##
## Coefficients:
##                            Estimate Std. Error  t value Pr(>|t|)
## (Intercept)              3.890e+00  3.457e-03 1125.258  < 2e-16 ***
## Hours_Studied            4.365e-03  7.339e-05   59.474  < 2e-16 ***
## Attendance               2.946e-03  3.808e-05   77.365  < 2e-16 ***
## Peer_InfluenceNeutral   -7.629e-03  9.885e-04   -7.717 1.36e-14 ***
## Peer_InfluenceNegative -1.535e-02  1.188e-03  -12.921  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03573 on 6602 degrees of freedom
## Multiple R-squared:  0.5928, Adjusted R-squared:  0.5926
```

```
## F-statistic:  2403 on 4 and 6602 DF,  p-value: < 2.2e-16
```

```r
# Plot Residuals vs. Fitted Values for the transformed model
plot(model_transformed$residuals ~ model_transformed$fitted.values,
     main = "Residuals vs Fitted (Transformed Model)",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")
```




```r
# Recheck Cook's Distance for influential points
cooksd_transformed <- cooks.distance(model_transformed)
plot(cooksd_transformed, main = "Cook's Distance (Transformed Model)", ylab = "Cook's Distance")
abline(h = 4/length(cooksd_transformed), col = "red")
```




```
# DW test for the transformed model
library(lmtest)
dwtest(model_transformed)
```

```
##
##  Durbin-Watson test
##
## data:  model_transformed
## DW = 2.0008, p-value = 0.5132
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# Check VIF for the transformed model
vif(model_transformed)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## Hours_Studied  1.000232  1        1.000116
## Attendance     1.000888  1        1.000444
## Peer_Influence 1.000928  2        1.000232
```