Regresyon Projesi

2026-01-02

1-) İlk olarak Gerekli Paketleri Aktif Edelim

```
library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)
library(ISLR)
library(caret)
library(mice)
library(PerformanceAnalytics)
library(car)
library(lmtest)
library(nortest)
library(readr)
```

2-) Data'ya İlk Bakışı Yapalım

- Data'yı ilk olarak df değeri olarak atayalım.

```
df <- read_csv("C:/Users/asus/Downloads/student_habits_performance.csv")
```

```
## Rows: 1000 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): student_id, gender, part_time_job, diet_quality, parental_education...
## dbl (9): age, study_hours_per_day, social_media_hours, netflix_hours, attend...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

- Ardından df'deki değişkenleri inceleyelim.

```
names(df)
```

```
##  [1] "student_id"                    "age"
##  [3] "gender"                        "study_hours_per_day"
##  [5] "social_media_hours"            "netflix_hours"
##  [7] "part_time_job"                 "attendance_percentage"
##  [9] "sleep_hours"                   "diet_quality"
## [11] "exercise_frequency"            "parental_education_level"
## [13] "internet_quality"              "mental_health_rating"
## [15] "extracurricular_participation" "exam_score"
```

- df'ye genel bi bakış atalım ve head() komutu ile küçük bir kısmını inceleyelim.

```
str(df)
```

```
## spc_tbl_ [1,000 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ student_id                   : chr [1:1000] "S1000" "S1001" "S1002" "S1003" ...
##  $ age                          : num [1:1000] 23 20 21 23 19 24 21 21 23 18 ...
##  $ gender                       : chr [1:1000] "Female" "Female" "Male" "Female" ...
##  $ study_hours_per_day          : num [1:1000] 0 6.9 1.4 1 5 7.2 5.6 4.3 4.4 4.8 ...
##  $ social_media_hours           : num [1:1000] 1.2 2.8 3.1 3.9 4.4 1.3 1.5 1 2.2 3.1 ...
##  $ netflix_hours                : num [1:1000] 1.1 2.3 1.3 1 0.5 0 1.4 2 1.7 1.3 ...
##  $ part_time_job                : chr [1:1000] "No" "No" "No" "No" ...
##  $ attendance_percentage        : num [1:1000] 85 97.3 94.8 71 90.9 82.9 85.8 77.7 100 95.4 ...
##  $ sleep_hours                  : num [1:1000] 8 4.6 8 9.2 4.9 7.4 6.5 4.6 7.1 7.5 ...
##  $ diet_quality                 : chr [1:1000] "Fair" "Good" "Poor" "Poor" ...
##  $ exercise_frequency           : num [1:1000] 6 6 1 4 3 1 2 0 3 5 ...
##  $ parental_education_level     : chr [1:1000] "Master" "High School" "High School" "Master" ...
##  $ internet_quality             : chr [1:1000] "Average" "Average" "Poor" "Good" ...
##  $ mental_health_rating         : num [1:1000] 8 8 1 1 1 4 4 8 1 10 ...
##  $ extracurricular_participation: chr [1:1000] "Yes" "No" "No" "Yes" ...
##  $ exam_score                   : num [1:1000] 56.2 100 34.3 26.8 66.4 100 89.8 72.6 78.9 100 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   student_id = col_character(),
##   ..   age = col_double(),
##   ..   gender = col_character(),
##   ..   study_hours_per_day = col_double(),
##   ..   social_media_hours = col_double(),
##   ..   netflix_hours = col_double(),
##   ..   part_time_job = col_character(),
##   ..   attendance_percentage = col_double(),
##   ..   sleep_hours = col_double(),
##   ..   diet_quality = col_character(),
##   ..   exercise_frequency = col_double(),
##   ..   parental_education_level = col_character(),
##   ..   internet_quality = col_character(),
##   ..   mental_health_rating = col_double(),
##   ..   extracurricular_participation = col_character(),
##   ..   exam_score = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(df)
```

```
## # A tibble: 6 x 16
##   student_id   age gender study_hours_per_day social_media_hours netflix_hours
##   <chr>      <dbl> <chr>                <dbl>              <dbl>         <dbl>
## 1 S1000         23 Female                 0                  1.2           1.1
## 2 S1001         20 Female                 6.9                2.8           2.3
## 3 S1002         21 Male                   1.4                3.1           1.3
## 4 S1003         23 Female                 1                  3.9           1
## 5 S1004         19 Female                 5                  4.4           0.5
## 6 S1005         24 Male                   7.2                1.3           0
```

```
## # i 10 more variables: part_time_job <chr>, attendance_percentage <dbl>,
## #   sleep_hours <dbl>, diet_quality <chr>, exercise_frequency <dbl>,
## #   parental_education_level <chr>, internet_quality <chr>,
## #   mental_health_rating <dbl>, extracurricular_participation <chr>,
## #   exam_score <dbl>
```

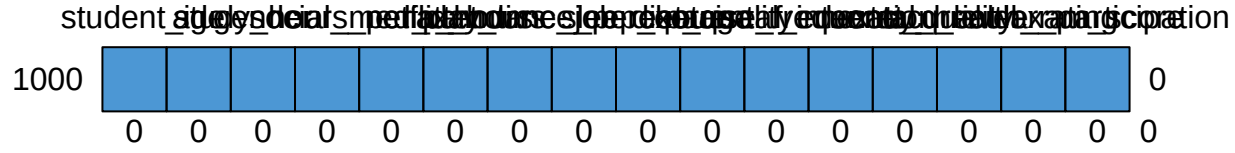
3-) Eksik Gözlemleri İnceleyelim

```
sum(is.na(df))
```

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

```
md.pattern(df)
```

```
##  /\     /\
## {  `---'  }
## {  0   0  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```




```
##      student_id age gender study_hours_per_day social_media_hours netflix_hours
```

```
## 1000          1   1      1                   1                  1             1
##               0   0      0                   0                  0             0
##      part_time_job attendance_percentage sleep_hours diet_quality
## 1000             1                    1           1            1
##                  0                    0           0            0
##      exercise_frequency parental_education_level internet_quality
## 1000                  1                        1                1
##                       0                        0                0
##      mental_health_rating extracurricular_participation exam_score
## 1000                    1                             1          1 0
##                         0                             0          0 0
```

- df'mizde eksik gözlem yok. Fakat bu bilgimizi pekiştirmek adına kendimiz rastgele eksik veri atayalım ve bunları dolduralım.

```
datanumeric <- Filter(is.numeric,df)

eksik_gozlem <- datanumeric[c("study_hours_per_day", "social_media_hours")]

head(eksik_gozlem)
```

```
## # A tibble: 6 x 2
##   study_hours_per_day social_media_hours
##                 <dbl>              <dbl>
## 1                   0                1.2
## 2                 6.9                2.8
## 3                 1.4                3.1
## 4                   1                3.9
## 5                   5                4.4
## 6                 7.2                1.3
```

```
for(i in colnames(eksik_gozlem)){
  idx <- sample(seq_len(nrow(datanumeric)),size = 10)
  datanumeric[idx,i] <- NA
}
sum(is.na(datanumeric))
```

```
## [1] 20
```

```
df["study_hours_per_day"] <- datanumeric["study_hours_per_day"]
df["social_media_hours"] <- datanumeric["social_media_hours"]
sum(is.na(df))
```

```
## [1] 20
```

- Burada sadece numeric değerlerin olduğu datanumeric isimli bi dataframe oluşturduk ve bu dataframe içerisinden 2 değişkene toplam 20 tane rastgele eksik gözlem atamak adına bir döngü yaptık. Ardından bu eksik gözlemleri df değerine atadık. Toplam 20 eksik gözlem oldu.
- Şimdi bu eksik gözlemleri dolduralım;

```r
set.seed(123)
imputed <- mice(df, m = 3)
```

```
##
##  iter imp variable
##   1   1  study_hours_per_day  social_media_hours
##   1   2  study_hours_per_day  social_media_hours
##   1   3  study_hours_per_day  social_media_hours
##   2   1  study_hours_per_day  social_media_hours
##   2   2  study_hours_per_day  social_media_hours
##   2   3  study_hours_per_day  social_media_hours
##   3   1  study_hours_per_day  social_media_hours
##   3   2  study_hours_per_day  social_media_hours
##   3   3  study_hours_per_day  social_media_hours
##   4   1  study_hours_per_day  social_media_hours
##   4   2  study_hours_per_day  social_media_hours
##   4   3  study_hours_per_day  social_media_hours
##   5   1  study_hours_per_day  social_media_hours
##   5   2  study_hours_per_day  social_media_hours
##   5   3  study_hours_per_day  social_media_hours
```

```
## Warning: Number of logged events: 7
```

```r
imputed$imp
```

```
## $student_id
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $age
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $gender
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $study_hours_per_day
##       1   2   3
## 28  5.5 4.9 4.3
## 63  5.6 5.9 7.4
## 160 3.6 3.4 4.0
## 418 6.3 5.4 6.3
## 473 2.3 2.4 2.9
## 483 3.3 3.3 4.0
## 500 5.8 5.1 4.5
## 574 2.2 3.0 2.6
## 716 5.2 4.5 5.8
## 861 3.3 3.2 3.8
##
## $social_media_hours
##       1   2   3
```

```
## 11  4.4 4.3 2.9
## 80  3.8 3.5 1.2
## 111 3.6 0.7 2.6
## 414 0.8 2.3 2.4
## 578 2.1 2.1 2.5
## 629 0.0 1.5 1.4
## 726 0.8 1.8 2.5
## 743 2.7 2.8 2.4
## 752 0.6 1.9 2.8
## 881 3.2 1.5 1.6
##
## $netflix_hours
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $part_time_job
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $attendance_percentage
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $sleep_hours
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $diet_quality
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $exercise_frequency
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $parental_education_level
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $internet_quality
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $mental_health_rating
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $extracurricular_participation
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
##
## $exam_score
## [1] 1 2 3
## <0 rows> (or 0-length row.names)
```

```
df <- complete(imputed,3)
sum(is.na(df))
```

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

```
md.pattern(df)
```

```
##  /\     /\
## {  `---'  }
## {  0   0  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```


```
##      student_id age gender study_hours_per_day social_media_hours netflix_hours
## 1000          1   1      1                   1                   1             1
##               0   0      0                   0                   0             0
##      part_time_job attendance_percentage sleep_hours diet_quality
## 1000             1                     1           1            1
##                  0                     0           0            0
##      exercise_frequency parental_education_level internet_quality
## 1000                  1                       1                1
##                       0                       0                0
##      mental_health_rating extracurricular_participation exam_score
```

```
## 1000                    1                              1          1 0
##                         0                              0          0 0
```

- Tabi oluşturduğumuz datanumeric dataframe'ini de doldurmalıyız. Bunun için direkt olarak doldurduğumuz df değerinden tekrar çekme işlemi yapabiliriz.

```
datanumeric <- Filter(is.numeric,df)
sum(is.na(datanumeric))
```

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

- datanumeric'in korelasyonuna bakalım;

```
cor(datanumeric)
```

```
##                               age study_hours_per_day social_media_hours
## age                    1.000000000        0.0031347581      -0.0102464342
## study_hours_per_day    0.003134758        1.0000000000       0.0195086559
## social_media_hours    -0.010246434        0.0195086559       1.0000000000
## netflix_hours         -0.001174104       -0.0279724507       0.0114077316
## attendance_percentage -0.026055201        0.0294715055       0.0429916018
## sleep_hours            0.037481916       -0.0266401965       0.0221844101
## exercise_frequency    -0.003836236       -0.0256847050      -0.0355093092
## mental_health_rating  -0.045101361       -0.0006814067       0.0001947846
## exam_score            -0.008906872        0.8260270424      -0.1665989877
##                       netflix_hours attendance_percentage   sleep_hours
## age                   -0.0011741040          -0.026055201  0.0374819156
## study_hours_per_day   -0.0279724507           0.029471506 -0.0266401965
## social_media_hours     0.0114077316           0.042991602  0.0221844101
## netflix_hours          1.0000000000          -0.002091540 -0.0009345491
## attendance_percentage -0.0020915397           1.000000000  0.0137560647
## sleep_hours           -0.0009345491           0.013756065  1.0000000000
## exercise_frequency    -0.0064482222          -0.007857196  0.0197690236
## mental_health_rating   0.0080342346          -0.018744560 -0.0065079649
## exam_score            -0.1717792385           0.089835602  0.1216829106
##                       exercise_frequency mental_health_rating   exam_score
## age                       -0.0038362359        -0.0451013606 -0.008906872
## study_hours_per_day       -0.0256847050        -0.0006814067  0.826027042
## social_media_hours        -0.0355093092         0.0001947846 -0.166598988
## netflix_hours             -0.0064482222         0.0080342346 -0.171779238
## attendance_percentage     -0.0078571964        -0.0187445601  0.089835602
## sleep_hours                0.0197690236        -0.0065079649  0.121682911
## exercise_frequency         1.0000000000        -0.0002422927  0.160107464
## mental_health_rating      -0.0002422927         1.0000000000  0.321522931
## exam_score                 0.1601074644         0.3215229307  1.000000000
```

4-) Eğitim Test Parçalanması Yapalım Ve Aykırı Gözlemleri İnceleyelim

```
sum(sapply(datanumeric,is.numeric))
```

```
## [1] 9
```

```
set.seed(123)

sampleind <- sample(1:nrow(df), size = 0.8*nrow(df))
trainset <- df[sampleind,]
testset <- df[-sampleind,]

train_dis <- datanumeric[sampleind,]
nrow(train_dis)
```

```
## [1] 800
```

```
nrow(trainset);nrow(testset)
```

```
## [1] 800

## [1] 200
```

```
distance <- mahalanobis(train_dis,center = colMeans(train_dis),cov = cov(train_dis))
cutoff <- qchisq(p = 0.95, df = 16)
ids <- which(distance > cutoff)

trainsetrem <- trainset[-ids,]
nrow(trainset);nrow(trainsetrem)
```

```
## [1] 800

## [1] 796
```

- Mahalanobis sayısal değerlere uygulanabildiği için datanumeric üzerinden mahalanobis uyguladık ve çıkan aykırıların indislerini df ile oluşturduğumuz trainsetten temizledik. Artık hem aykırısız hem de aykırılı iki trainsetimiz var.

5-) Artık Faktör Atamalarını Yapabiliriz.

```
str(trainset)
```

```
## 'data.frame':    800 obs. of  16 variables:
##  $ student_id                : chr  "S1414" "S1462" "S1178" "S1525" ...
##  $ age                       : num  19 22 17 23 23 21 19 22 21 17 ...
##  $ gender                    : chr  "Female" "Male" "Other" "Male" ...
##  $ study_hours_per_day       : num  4.2 4.9 2.9 5.3 2.6 1.9 4.9 4.1 5.4 3.5 ...
##  $ social_media_hours        : num  5.6 2.6 2.5 2.9 4 3.2 1.5 1.1 3.1 3.6 ...
##  $ netflix_hours             : num  0.4 0 2.4 0.9 1.1 1.2 3.4 1.9 1.8 2.7 ...
##  $ part_time_job             : chr  "No" "Yes" "Yes" "No" ...
##  $ attendance_percentage     : num  91.1 66.1 83.7 91.6 89.4 75.9 71.4 82.3 88.1 92.9 ...
##  $ sleep_hours               : num  9.8 7.6 6.3 5 6.9 6 6.9 7.5 6.2 8.4 ...
##  $ diet_quality              : chr  "Poor" "Poor" "Poor" "Good" ...
```

```
##  $ exercise_frequency           : num  0 3 5 4 0 6 6 1 3 3 ...
##  $ parental_education_level     : chr  "Bachelor" "Bachelor" "None" "High School" ...
##  $ internet_quality             : chr  "Poor" "Good" "Good" "Poor" ...
##  $ mental_health_rating         : num  5 4 9 4 10 5 7 8 4 8 ...
##  $ extracurricular_participation: chr  "Yes" "No" "No" "No" ...
##  $ exam_score                   : num  78.7 80.9 65.2 84.1 66.6 41.7 88.2 80.8 91.3 72.6 ...
```

```
str(trainsetrem)
```

```
## 'data.frame':    796 obs. of  16 variables:
##  $ student_id                   : chr  "S1414" "S1462" "S1178" "S1525" ...
##  $ age                          : num  19 22 17 23 23 21 19 22 21 17 ...
##  $ gender                       : chr  "Female" "Male" "Other" "Male" ...
##  $ study_hours_per_day          : num  4.2 4.9 2.9 5.3 2.6 1.9 4.9 4.1 5.4 3.5 ...
##  $ social_media_hours           : num  5.6 2.6 2.5 2.9 4 3.2 1.5 1.1 3.1 3.6 ...
##  $ netflix_hours                : num  0.4 0 2.4 0.9 1.1 1.2 3.4 1.9 1.8 2.7 ...
##  $ part_time_job                : chr  "No" "Yes" "Yes" "No" ...
##  $ attendance_percentage        : num  91.1 66.1 83.7 91.6 89.4 75.9 71.4 82.3 88.1 92.9 ...
##  $ sleep_hours                  : num  9.8 7.6 6.3 5 6.9 6 6.9 7.5 6.2 8.4 ...
##  $ diet_quality                 : chr  "Poor" "Poor" "Poor" "Good" ...
##  $ exercise_frequency           : num  0 3 5 4 0 6 6 1 3 3 ...
##  $ parental_education_level     : chr  "Bachelor" "Bachelor" "None" "High School" ...
##  $ internet_quality             : chr  "Poor" "Good" "Good" "Poor" ...
##  $ mental_health_rating         : num  5 4 9 4 10 5 7 8 4 8 ...
##  $ extracurricular_participation: chr  "Yes" "No" "No" "No" ...
##  $ exam_score                   : num  78.7 80.9 65.2 84.1 66.6 41.7 88.2 80.8 91.3 72.6 ...
```

```
trainset$gender <- as.factor(trainset$gender)
trainset$part_time_job <- as.factor(trainset$part_time_job)
trainset$diet_quality <- as.factor(trainset$diet_quality)
trainset$parental_education_level <- as.factor(trainset$parental_education_level)
trainset$internet_quality <- as.factor(trainset$internet_quality)
trainset$extracurricular_participation <- as.factor(trainset$extracurricular_participation)
trainset$student_id <- NULL


trainsetrem$gender <- as.factor(trainsetrem$gender)
trainsetrem$part_time_job <- as.factor(trainsetrem$part_time_job)
trainsetrem$diet_quality <- as.factor(trainsetrem$diet_quality)
trainsetrem$parental_education_level <- as.factor(trainsetrem$parental_education_level)
trainsetrem$internet_quality <- as.factor(trainsetrem$internet_quality)
trainsetrem$extracurricular_participation <- as.factor(trainsetrem$extracurricular_participation)
trainsetrem$student_id <- NULL
```

- Hem aykırısız hem de aykırılı trainsetlerimize faktör atamalarını yaptık ve student_id değişkenini çıkardık.

6-) İlk Modelleri Oluşturalım

```
set.seed(123)
trainset_model1 <- lm(exam_score~.,data = trainset)
summary(trainset_model1)
```

```
##
## Call:
## lm(formula = exam_score ~ ., data = trainset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2327  -3.4213  -0.0636   3.5313  16.1878
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         8.91887    2.84876   3.131  0.00181 **
## age                                -0.01766    0.08358  -0.211  0.83274
## genderMale                         -0.10590    0.39350  -0.269  0.78791
## genderOther                         1.28326    0.93364   1.374  0.16969
## study_hours_per_day                 9.41593    0.13129  71.719  < 2e-16 ***
## social_media_hours                 -2.54329    0.16813 -15.127  < 2e-16 ***
## netflix_hours                      -2.21557    0.17698 -12.519  < 2e-16 ***
## part_time_jobYes                   -0.06733    0.46790  -0.144  0.88561
## attendance_percentage               0.13342    0.02089   6.387 2.91e-10 ***
## sleep_hours                         1.90516    0.15539  12.261  < 2e-16 ***
## diet_qualityGood                   -0.97150    0.42710  -2.275  0.02320 *
## diet_qualityPoor                   -0.20490    0.53267  -0.385  0.70058
## exercise_frequency                  1.46183    0.09389  15.570  < 2e-16 ***
## parental_education_levelHigh School  0.15760    0.44862   0.351  0.72546
## parental_education_levelMaster     -0.22712    0.57106  -0.398  0.69095
## parental_education_levelNone       -0.80586    0.71363  -1.129  0.25915
## internet_qualityGood               -0.56862    0.42153  -1.349  0.17774
## internet_qualityPoor                0.14592    0.57705   0.253  0.80044
## mental_health_rating                1.89109    0.06805  27.792  < 2e-16 ***
## extracurricular_participationYes    0.54979    0.40812   1.347  0.17832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.381 on 780 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8947
## F-statistic: 358.2 on 19 and 780 DF,  p-value: < 2.2e-16
```

```
vif(trainset_model1)
```

```
##                                  GVIF Df GVIF^(1/(2*Df))
## age                          1.019410  1        1.009658
## gender                       1.037484  2        1.009242
## study_hours_per_day          1.018400  1        1.009158
## social_media_hours           1.018533  1        1.009224
## netflix_hours                1.012117  1        1.006040
## part_time_job                1.016699  1        1.008315
## attendance_percentage        1.027202  1        1.013510
## sleep_hours                  1.012869  1        1.006414
## diet_quality                 1.043000  2        1.010581
## exercise_frequency           1.015268  1        1.007605
## parental_education_level     1.060647  3        1.009862
## internet_quality             1.049573  2        1.012169
## mental_health_rating         1.033837  1        1.016778
## extracurricular_participation 1.013669  1        1.006811
```

-Aykırılı olan modelimizin anlamlı olduğunu ve modeldeki değişkenlerin bağımlı değişkenimiz olan exam_score'u %89.68 oranında açıklayabildiğini görüyoruz. vif değerleri ise 1'e çok yakın yani multicollinearity tehdidi yok.

-Şimdi aykırısız olan modelimize bakalım;

```
set.seed(123)
trainsetrem_model1 <- lm(exam_score~.,data = trainsetrem)
summary(trainsetrem_model1)
```

```
##
## Call:
## lm(formula = exam_score ~ ., data = trainsetrem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9375  -3.4723  -0.0778   3.3950  15.9002
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         6.48511    2.76406   2.346   0.0192 *
## age                                -0.02505    0.08049  -0.311   0.7557
## genderMale                         -0.09664    0.37952  -0.255   0.7991
## genderOther                         1.10294    0.89849   1.228   0.2200
## study_hours_per_day                 9.62384    0.12887  74.680  < 2e-16 ***
## social_media_hours                 -2.50863    0.16200 -15.485  < 2e-16 ***
## netflix_hours                      -2.22335    0.17086 -13.013  < 2e-16 ***
## part_time_jobYes                    0.15701    0.45235   0.347   0.7286
## attendance_percentage               0.14562    0.02017   7.220 1.24e-12 ***
## sleep_hours                         1.93724    0.14970  12.940  < 2e-16 ***
## diet_qualityGood                   -0.96584    0.41217  -2.343   0.0194 *
## diet_qualityPoor                   -0.29806    0.51281  -0.581   0.5612
## exercise_frequency                  1.53292    0.09075  16.891  < 2e-16 ***
## parental_education_levelHigh School  0.28450    0.43288   0.657   0.5112
## parental_education_levelMaster      -0.23577    0.54961  -0.429   0.6681
## parental_education_levelNone        -0.84469    0.68679  -1.230   0.2191
## internet_qualityGood                -0.43372    0.40665  -1.067   0.2865
## internet_qualityPoor                 0.08995    0.55532   0.162   0.8714
## mental_health_rating                 1.94858    0.06584  29.595  < 2e-16 ***
## extracurricular_participationYes     0.55113    0.39349   1.401   0.1617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.176 on 776 degrees of freedom
## Multiple R-squared:  0.9038, Adjusted R-squared:  0.9014
## F-statistic: 383.7 on 19 and 776 DF,  p-value: < 2.2e-16
```

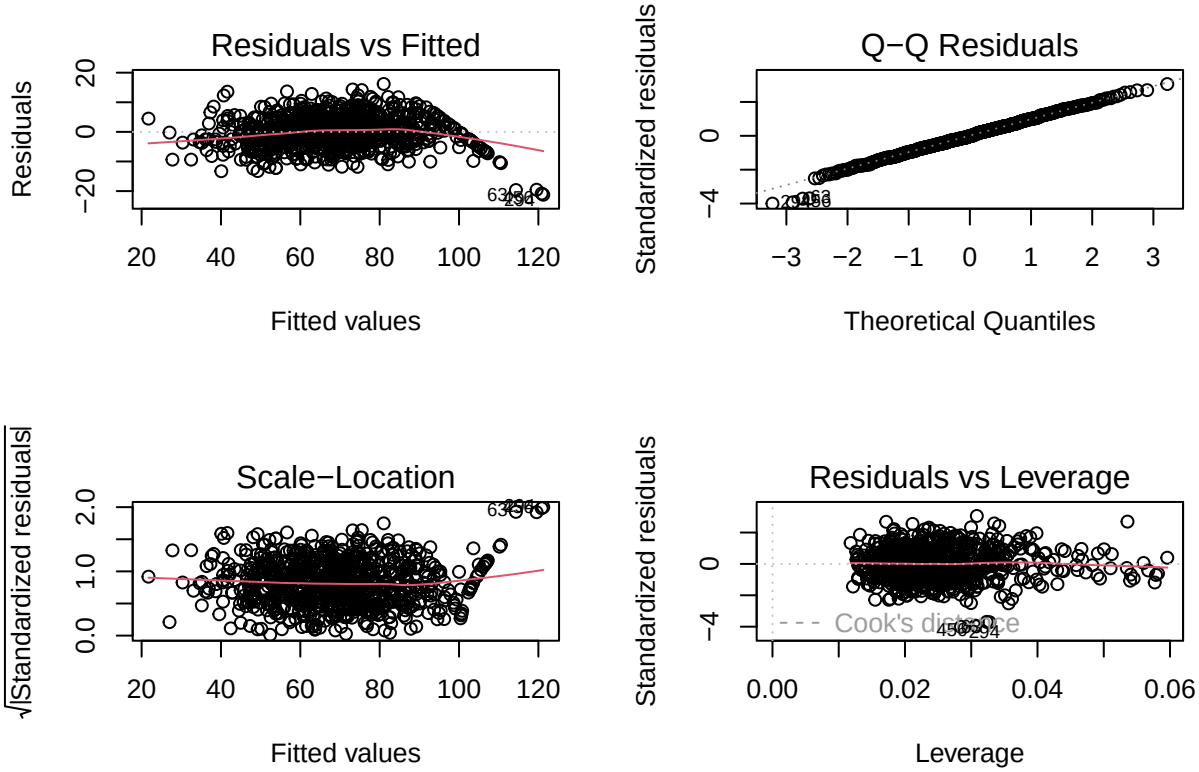
```
vif(trainsetrem_model1)
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## age                        1.019320  1        1.009614
## gender                     1.037870  2        1.009336
## study_hours_per_day        1.019672  1        1.009788
```

```
## social_media_hours            1.018168  1        1.009043
## netflix_hours                 1.012205  1        1.006084
## part_time_job                 1.016636  1        1.008284
## attendance_percentage         1.027312  1        1.013564
## sleep_hours                   1.013537  1        1.006746
## diet_quality                  1.042529  2        1.010467
## exercise_frequency            1.017283  1        1.008605
## parental_education_level      1.059122  3        1.009619
## internet_quality              1.049691  2        1.012198
## mental_health_rating          1.034469  1        1.017089
## extracurricular_participation 1.013712  1        1.006833
```

- Yine anlamlı ve daha yüksek R^2 değerine sahip bir model görüyoruz. Değişkenlerimiz, exam_score bağımlı değişkenimizi %90.21 oranında açıklayabiliyor.
- Tabi ki iki modelde anlamsız değişkenlerimiz var ve onları çıkaracağız.
- Fakat bundan önce ilk modellerimiz varsayımları sağlıyor mu kontrol edelim.

```
par(mfrow = c(2,2))
plot(trainset_model1)
```



- Aykırılı modelimizde normallik eğrisinde uç değerler biraz sıkıntılı duruyor.
- İlk grafikte de acaba değişen varyans durumu var mı diye sorgulatıyor.
- Dördüncü grafikten de kaldıraç değerlerimizin olması mümkün duruyor.

- Emin olmak için testleri uygulayalım.

```
bptest(trainset_model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainset_model1
## BP = 23.52, df = 19, p-value = 0.2152
```

```
shapiro.test(residuals(trainset_model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainset_model1)
## W = 0.99478, p-value = 0.007628
```

```
dwtest(trainset_model1)
```

```
##
##  Durbin-Watson test
##
## data:  trainset_model1
## DW = 1.9136, p-value = 0.1113
## alternative hypothesis: true autocorrelation is greater than 0
```

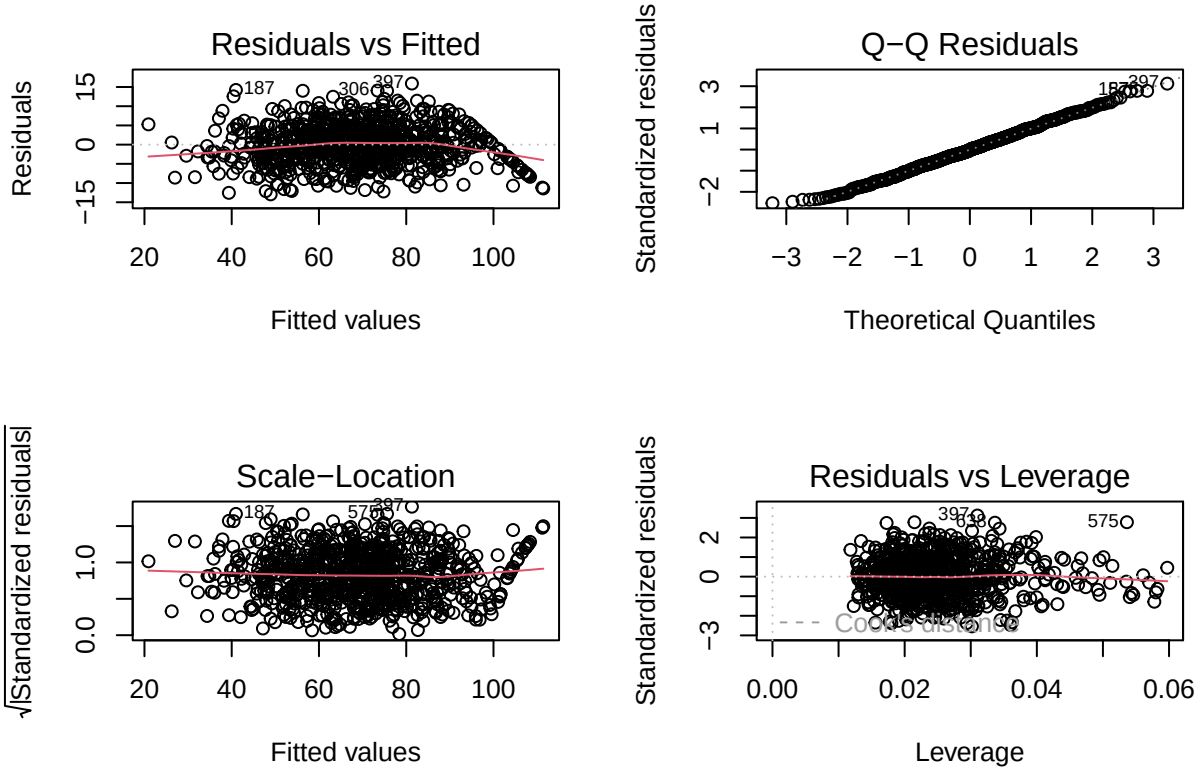
```
vif(trainset_model1)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## age                            1.019410  1        1.009658
## gender                         1.037484  2        1.009242
## study_hours_per_day            1.018400  1        1.009158
## social_media_hours             1.018533  1        1.009224
## netflix_hours                  1.012117  1        1.006040
## part_time_job                  1.016699  1        1.008315
## attendance_percentage          1.027202  1        1.013510
## sleep_hours                    1.012869  1        1.006414
## diet_quality                   1.043000  2        1.010581
## exercise_frequency             1.015268  1        1.007605
## parental_education_level       1.060647  3        1.009862
## internet_quality               1.049573  2        1.012169
## mental_health_rating           1.033837  1        1.016778
## extracurricular_participation 1.013669  1        1.006811
```

- bptest() ile sabit varyans olduğunu gördük.
- Shapiro ile normallik varsayımının sağlanmadığını gördük.
- dwtest ile otokorelasyon olmadığını gördük.
- vif ile de multicollinearity olmadığını gördük.

- Yani aykrılı modelimizin normallik varsayımını sağlamadığını anladık.
- Şimdi aykrısız modeli inceleyelim.

```
par(mfrow = c(2,2))
plot(trainsetrem_model1)
```



- Normallik eğrisinin aykırılı modele göre daha iyi olduğu gözüküyor.
- Yine değişen varyans sorunu var mı anlamak için test uygulayacağız.
- Kaldıraç değerlerimiz de aynı şekilde olabilir.

```
bptest(trainsetrem_model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model1
## BP = 11.873, df = 19, p-value = 0.891
```

```
shapiro.test(residuals(trainsetrem_model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model1)
## W = 0.9978, p-value = 0.3845
```

```
dwtest(trainsetrem_model1)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model1
## DW = 1.9035, p-value = 0.08703
## alternative hypothesis: true autocorrelation is greater than 0
```

```
vif(trainsetrem_model1)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## age                            1.019320  1        1.009614
## gender                         1.037870  2        1.009336
## study_hours_per_day            1.019672  1        1.009788
## social_media_hours             1.018168  1        1.009043
## netflix_hours                  1.012205  1        1.006084
## part_time_job                  1.016636  1        1.008284
## attendance_percentage          1.027312  1        1.013564
## sleep_hours                    1.013537  1        1.006746
## diet_quality                   1.042529  2        1.010467
## exercise_frequency             1.017283  1        1.008605
## parental_education_level       1.059122  3        1.009619
## internet_quality               1.049691  2        1.012198
## mental_health_rating           1.034469  1        1.017089
## extracurricular_participation 1.013712  1        1.006833
```

- Değişen varyans sorunu olmadığını ve shapiro testi ilenormalliğin artık sağlandığını anladık. Otokorelasyon ve multicollinearity gibi sorunların da olmadığını görmüş olduk aykırısız modelde.
- Bu nedenle buradan itibaren sadece aykırısız model ile yola devam edeceğiz. Çünkü değerler olarak hem daha iyi hem de varsayımları sağlıyor.

7-) Anlamsız değişkenleri çıkaralım;

- Oluşturduğumuz her yeni modelde tekrar varsayım kontrollerini yapıyoruz.

```
trainsetrem_model2 <- lm(exam_score~. -internet_quality, data = trainsetrem)
summary(trainsetrem_model2)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality, data = trainsetrem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7677  -3.4163  -0.1178   3.3231  16.0847
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       6.36415    2.76138   2.305   0.0214 *
```

```
## age                                   -0.02934    0.08035  -0.365   0.7151
## genderMale                            -0.08147    0.37855  -0.215   0.8297
## genderOther                            1.07157    0.89770   1.194   0.2330
## study_hours_per_day                    9.62299    0.12869  74.775  < 2e-16 ***
## social_media_hours                    -2.51931    0.16173 -15.577  < 2e-16 ***
## netflix_hours                         -2.23327    0.17062 -13.089  < 2e-16 ***
## part_time_jobYes                       0.17785    0.45085   0.394   0.6933
## attendance_percentage                  0.14600    0.02013   7.255 9.75e-13 ***
## sleep_hours                            1.93510    0.14965  12.931  < 2e-16 ***
## diet_qualityGood                      -0.99463    0.41137  -2.418   0.0158 *
## diet_qualityPoor                      -0.26372    0.51176  -0.515   0.6065
## exercise_frequency                     1.53684    0.09061  16.962  < 2e-16 ***
## parental_education_levelHigh School    0.30838    0.43227   0.713   0.4758
## parental_education_levelMaster        -0.24828    0.54920  -0.452   0.6513
## parental_education_levelNone          -0.82389    0.68638  -1.200   0.2304
## mental_health_rating                   1.95305    0.06572  29.717  < 2e-16 ***
## extracurricular_participationYes       0.55639    0.39291   1.416   0.1572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.175 on 778 degrees of freedom
## Multiple R-squared:  0.9036, Adjusted R-squared:  0.9015
## F-statistic:   429 on 17 and 778 DF,  p-value: < 2.2e-16
```

```
bptest(trainsetrem_model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model2
## BP = 11.539, df = 17, p-value = 0.8272
```

```
shapiro.test(residuals(trainsetrem_model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model2)
## W = 0.99785, p-value = 0.4068
```

```
dwtest(trainsetrem_model2)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model2
## DW = 1.9056, p-value = 0.092
## alternative hypothesis: true autocorrelation is greater than 0
```

```
trainsetrem_model3 <- lm(exam_score~. -internet_quality -gender, data = trainsetrem)
summary(trainsetrem_model3)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender, data = trainsetrem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8367  -3.4552  -0.1058   3.4298  16.0085
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         6.21718    2.75237   2.259   0.0242 *
## age                                -0.02858    0.08030  -0.356   0.7220
## study_hours_per_day                 9.62755    0.12860  74.866  < 2e-16 ***
## social_media_hours                 -2.51193    0.16146 -15.558  < 2e-16 ***
## netflix_hours                      -2.22003    0.17027 -13.039  < 2e-16 ***
## part_time_jobYes                    0.18322    0.45061   0.407   0.6844
## attendance_percentage               0.14569    0.02012   7.241 1.07e-12 ***
## sleep_hours                         1.94714    0.14930  13.042  < 2e-16 ***
## diet_qualityGood                   -0.99851    0.41125  -2.428   0.0154 *
## diet_qualityPoor                   -0.25424    0.51153  -0.497   0.6193
## exercise_frequency                  1.53598    0.09038  16.995  < 2e-16 ***
## parental_education_levelHigh School  0.34473    0.43091   0.800   0.4239
## parental_education_levelMaster      -0.22834    0.54869  -0.416   0.6774
## parental_education_levelNone        -0.76708    0.68447  -1.121   0.2628
## mental_health_rating                1.95506    0.06565  29.780  < 2e-16 ***
## extracurricular_participationYes    0.55455    0.39280   1.412   0.1584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.174 on 780 degrees of freedom
## Multiple R-squared:  0.9034, Adjusted R-squared:  0.9015
## F-statistic: 486.3 on 15 and 780 DF,  p-value: < 2.2e-16
```

```
bptest(trainsetrem_model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model3
## BP = 9.41, df = 15, p-value = 0.8551
```

```
shapiro.test(residuals(trainsetrem_model3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model3)
## W = 0.99796, p-value = 0.4584
```

```
dwtest(trainsetrem_model3)
```

```
##
```

```
##  Durbin-Watson test
##
## data:  trainsetrem_model3
## DW = 1.9101, p-value = 0.1023
## alternative hypothesis: true autocorrelation is greater than 0
```

```
trainsetrem_model4 <- lm(exam_score~. -internet_quality -gender -age, data = trainsetrem)
summary(trainsetrem_model4)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender - age,
##     data = trainsetrem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9250  -3.4608  -0.0782   3.4731  15.9177
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         5.60795    2.15405   2.603   0.0094 **
## study_hours_per_day                 9.62781    0.12852  74.911  < 2e-16 ***
## social_media_hours                 -2.50994    0.16127 -15.563  < 2e-16 ***
## netflix_hours                      -2.22226    0.17005 -13.068  < 2e-16 ***
## part_time_jobYes                    0.18105    0.45032   0.402   0.6878
## attendance_percentage               0.14597    0.02009   7.265 9.03e-13 ***
## sleep_hours                         1.94457    0.14904  13.047  < 2e-16 ***
## diet_qualityGood                   -0.99688    0.41100  -2.426   0.0155 *
## diet_qualityPoor                   -0.25306    0.51123  -0.495   0.6207
## exercise_frequency                  1.53648    0.09032  17.012  < 2e-16 ***
## parental_education_levelHigh School  0.35537    0.42963   0.827   0.4084
## parental_education_levelMaster      -0.22099    0.54799  -0.403   0.6869
## parental_education_levelNone        -0.76063    0.68385  -1.112   0.2664
## mental_health_rating                1.95648    0.06549  29.873  < 2e-16 ***
## extracurricular_participationYes    0.55510    0.39258   1.414   0.1578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.171 on 781 degrees of freedom
## Multiple R-squared:  0.9034, Adjusted R-squared:  0.9016
## F-statistic: 521.6 on 14 and 781 DF,  p-value: < 2.2e-16
```

```
bptest(trainsetrem_model4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model4
## BP = 7.3196, df = 14, p-value = 0.9217
```

```
shapiro.test(residuals(trainsetrem_model4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model4)
## W = 0.99803, p-value = 0.4904
```

```
dwtest(trainsetrem_model4)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model4
## DW = 1.9088, p-value = 0.09904
## alternative hypothesis: true autocorrelation is greater than 0
```

```
trainsetrem_model5 <- lm(exam_score~. -internet_quality -gender -age -diet_quality, data = trainsetrem)
summary(trainsetrem_model5)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender - age -
##     diet_quality, data = trainsetrem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.408  -3.350  -0.175   3.493  16.253
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        4.96280    2.14341   2.315   0.0208 *
## study_hours_per_day                9.64058    0.12843  75.066  < 2e-16 ***
## social_media_hours                -2.51769    0.16161 -15.578  < 2e-16 ***
## netflix_hours                     -2.22216    0.17045 -13.037  < 2e-16 ***
## part_time_jobYes                   0.15004    0.45124   0.332   0.7396
## attendance_percentage              0.14869    0.02010   7.396 3.61e-13 ***
## sleep_hours                        1.95213    0.14939  13.067  < 2e-16 ***
## exercise_frequency                 1.53099    0.09045  16.926  < 2e-16 ***
## parental_education_levelHigh School  0.34838    0.43074   0.809   0.4189
## parental_education_levelMaster    -0.18975    0.54927  -0.345   0.7298
## parental_education_levelNone      -0.73777    0.68551  -1.076   0.2822
## mental_health_rating               1.94735    0.06548  29.738  < 2e-16 ***
## extracurricular_participationYes   0.51912    0.39230   1.323   0.1861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.184 on 783 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9011
## F-statistic: 604.9 on 12 and 783 DF,  p-value: < 2.2e-16
```

```
bptest(trainsetrem_model5)
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model5
## BP = 6.4238, df = 12, p-value = 0.8932
```

```
shapiro.test(residuals(trainsetrem_model5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model5)
## W = 0.9981, p-value = 0.528
```

```
dwtest(trainsetrem_model5)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model5
## DW = 1.9128, p-value = 0.1095
## alternative hypothesis: true autocorrelation is greater than 0
```

```
trainsetrem_model6 <- lm(exam_score~. -internet_quality -gender -age -diet_quality -part_time_job, data
summary(trainsetrem_model6)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender - age -
##     diet_quality - part_time_job, data = trainsetrem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.298  -3.311  -0.210   3.478  16.377
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        5.01394    2.13667   2.347   0.0192 *
## study_hours_per_day                9.63803    0.12813  75.223  < 2e-16 ***
## social_media_hours                -2.51784    0.16152 -15.588  < 2e-16 ***
## netflix_hours                     -2.22113    0.17032 -13.041  < 2e-16 ***
## attendance_percentage              0.14848    0.02008   7.394 3.67e-13 ***
## sleep_hours                        1.95270    0.14930  13.079  < 2e-16 ***
## exercise_frequency                 1.53067    0.09040  16.933  < 2e-16 ***
## parental_education_levelHigh School  0.35602    0.42988   0.828   0.4078
## parental_education_levelMaster    -0.18748    0.54892  -0.342   0.7328
## parental_education_levelNone      -0.73688    0.68512  -1.076   0.2825
## mental_health_rating               1.94738    0.06545  29.755  < 2e-16 ***
## extracurricular_participationYes   0.51854    0.39207   1.323   0.1864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.181 on 784 degrees of freedom
```

```
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9012
## F-statistic: 660.6 on 11 and 784 DF,  p-value: < 2.2e-16
```

```
bptest(trainsetrem_model6)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model6
## BP = 5.1186, df = 11, p-value = 0.9253
```

```
shapiro.test(residuals(trainsetrem_model6))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model6)
## W = 0.99808, p-value = 0.518
```

```
dwtest(trainsetrem_model6)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model6
## DW = 1.9125, p-value = 0.1089
## alternative hypothesis: true autocorrelation is greater than 0
```

```
trainsetrem_model7 <- lm(exam_score~. -internet_quality -gender -age -diet_quality -part_time_job -parer
summary(trainsetrem_model7)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender - age -
##     diet_quality - part_time_job - parental_education_level,
##     data = trainsetrem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3825  -3.3992  -0.1709   3.5054  16.2089
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.09666    2.09843   2.429   0.0154 *
## study_hours_per_day             9.63964    0.12812  75.237  < 2e-16 ***
## social_media_hours             -2.50793    0.16138 -15.540  < 2e-16 ***
## netflix_hours                  -2.21404    0.17021 -13.008  < 2e-16 ***
## attendance_percentage           0.14677    0.01998   7.345 5.14e-13 ***
## sleep_hours                     1.95351    0.14913  13.099  < 2e-16 ***
## exercise_frequency              1.53242    0.09039  16.954  < 2e-16 ***
## mental_health_rating            1.95549    0.06491  30.127  < 2e-16 ***
```

```
## extracurricular_participationYes  0.52871    0.39172   1.350   0.1775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.181 on 787 degrees of freedom
## Multiple R-squared:  0.9022, Adjusted R-squared:  0.9012
## F-statistic: 907.9 on 8 and 787 DF,  p-value: < 2.2e-16
```

```
bptest(trainsetrem_model7)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model7
## BP = 3.4903, df = 8, p-value = 0.8999
```

```
shapiro.test(residuals(trainsetrem_model7))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model7)
## W = 0.99828, p-value = 0.6238
```

```
dwtest(trainsetrem_model7)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model7
## DW = 1.9137, p-value = 0.1118
## alternative hypothesis: true autocorrelation is greater than 0
```

```
trainsetrem_model8 <- lm(exam_score~. -internet_quality -gender -age -diet_quality -part_time_job -parer
summary(trainsetrem_model8)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender - age -
##     diet_quality - part_time_job - parental_education_level -
##     extracurricular_participation, data = trainsetrem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5681  -3.2895  -0.2003   3.6002  16.0430
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.28920    2.09467   2.525   0.0118 *
## study_hours_per_day    9.63669    0.12817  75.186  < 2e-16 ***
## social_media_hours    -2.50990    0.16146 -15.545  < 2e-16 ***
```

```
## netflix_hours         -2.21606    0.17029 -13.013  < 2e-16 ***
## attendance_percentage  0.14661    0.01999   7.334 5.57e-13 ***
## sleep_hours            1.95868    0.14916  13.132  < 2e-16 ***
## exercise_frequency     1.52780    0.09037  16.906  < 2e-16 ***
## mental_health_rating   1.95432    0.06494  30.096  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.184 on 788 degrees of freedom
## Multiple R-squared:  0.902,  Adjusted R-squared:  0.9011
## F-statistic:  1036 on 7 and 788 DF,  p-value: < 2.2e-16
```

```
bptest(trainsetrem_model8)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model8
## BP = 0.65846, df = 7, p-value = 0.9986
```

```
shapiro.test(residuals(trainsetrem_model8))
```

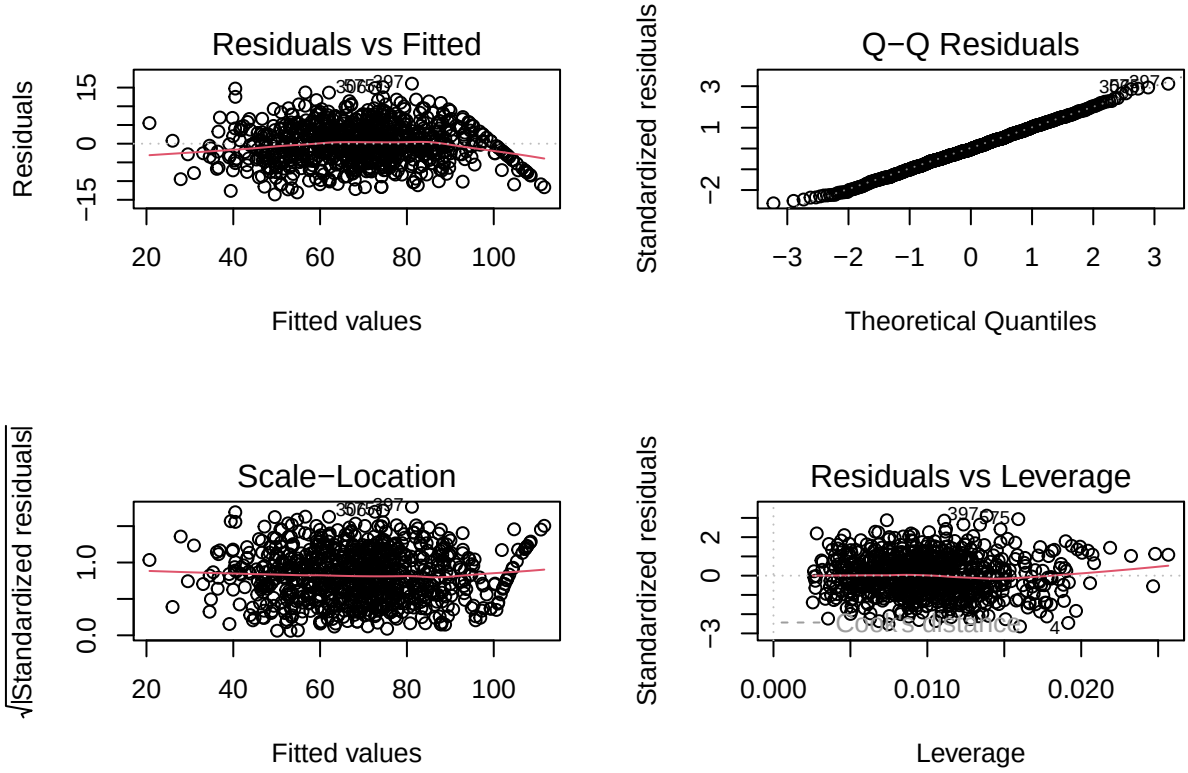
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model8)
## W = 0.99819, p-value = 0.5756
```

```
dwtest(trainsetrem_model8)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model8
## DW = 1.914, p-value = 0.1123
## alternative hypothesis: true autocorrelation is greater than 0
```

- Her adımda yaptığımız varsayım kontrolleri sonucunda bir sorun olmadığını görmüş olduk.
- Bütün değişkenleri anlamlı olan modelimizin varsayımlarını bir de grafikte görelim;

```
par(mfrow = c(2,2))
plot(trainsetrem_model8)
```



- Hiçbir sorun olmadığını grafikler ile de görmüş olduk.

- Anlamsızları çıkarmadığımız ve çıkardığımız modellerin AIC değerlerine bakalım

```
AIC(trainsetrem_model1)
```

```
## [1] 4898.067
```

```
AIC(trainsetrem_model8)
```

```
## [1] 4888.692
```

- Anlamsızları çıkardığımız modelin daha iyi olduğunu AIC ile gördük.

8-) Tahminleri Oluşturalım

```
tahminler_trainsetrem <- predict(trainsetrem_model8,testset)
sum(trainsetrem_model8$residuals)
```

```
## [1] -1.883771e-13
```

- Hataların toplamı 0'a çok yakın yani tahminlerde bir sorun yok.

9-) Modelimize Cooks Distance Ve Standartlaştırma Uygulayalım

```
strandardized_residuals_trainsetrem <- rstandard(trainsetrem_model8)
summary(strandardized_residuals_trainsetrem)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -2.638e+00 -6.363e-01 -3.887e-02 -2.944e-05  6.980e-01  3.116e+00
```

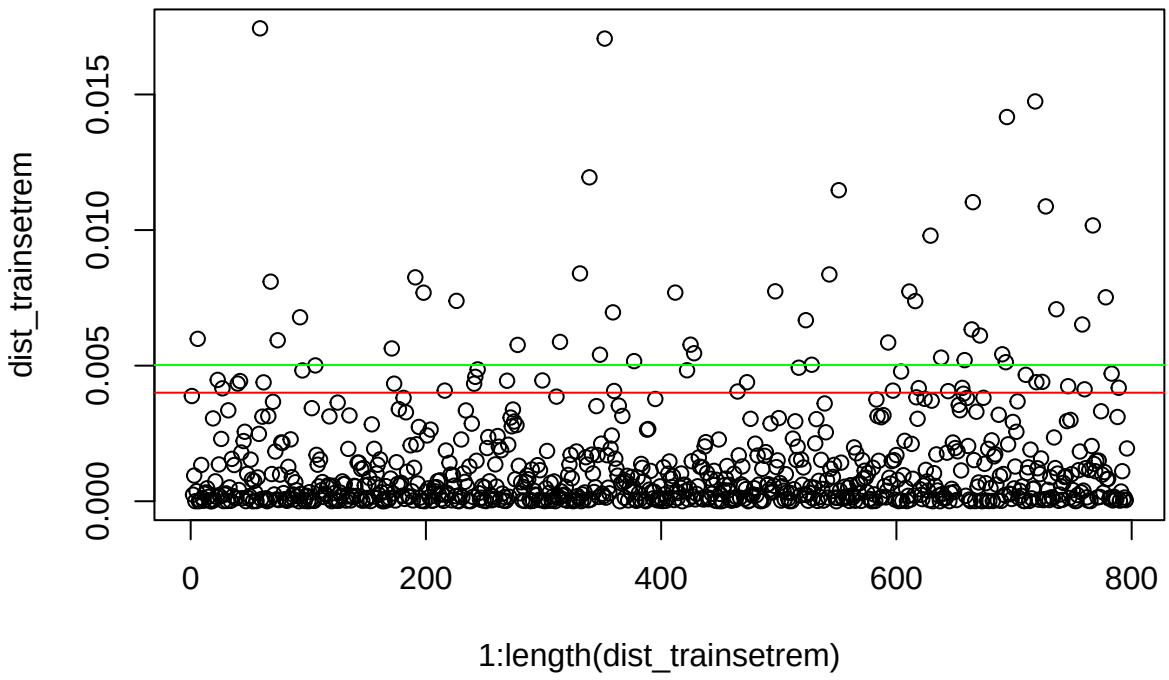
```
olcut1index <- which(abs(strandardized_residuals_trainsetrem) > 2)


dist_trainsetrem <- cooks.distance(trainsetrem_model8)
olcut1_rem <- mean(dist_trainsetrem) * 3
olcut2_rem <- 4/length(dist_trainsetrem)

olcut1index_rem <- which(dist_trainsetrem > olcut1_rem)
olcut2index_rem <- which(dist_trainsetrem > olcut2_rem)
length(olcut1index_rem);length(olcut2index_rem)
```

```
## [1] 74
```

```
## [1] 43
```

- Ölçütleri belirledik ve olçüt1'in diğerine göre daha fazla değer kapsadığını gördük. Bundan dolayı daha garanti olan ölçüt2'yi ana ölçütümüz olarak aldık.
- Grafikte de görelim;

```
plot(1:length(dist_trainsetrem),dist_trainsetrem,type = "p", ylim = range(dist_trainsetrem)*c(1,1))
abline(h = olcut1_rem, col = "red")
abline(h = olcut2_rem, col = "green")
```



- En büyük cook değerlerimizin bile 0.015 gibi düşük kaldığını görüyoruz. Yani bu değerleri atmasak da olur fakat her iki halini de inceleyelim yine de.
- Şimdi outlier olanları belirleyelim;

```
outliers_trainsetrem <- which(dist_trainsetrem > olcut2_rem & abs(strandardized_residuals_trainsetrem) >
length(outliers_trainsetrem)
```

```
## [1] 26
```

- 25 adet outlier değer bulduk.
- Şimdi bu değerleri de atalım;

```
trainsetrem_cooks_rem <- trainsetrem[-outliers_trainsetrem,]
nrow(trainsetrem);nrow(trainsetrem_cooks_rem)
```

```
## [1] 796
```

```
## [1] 770
```

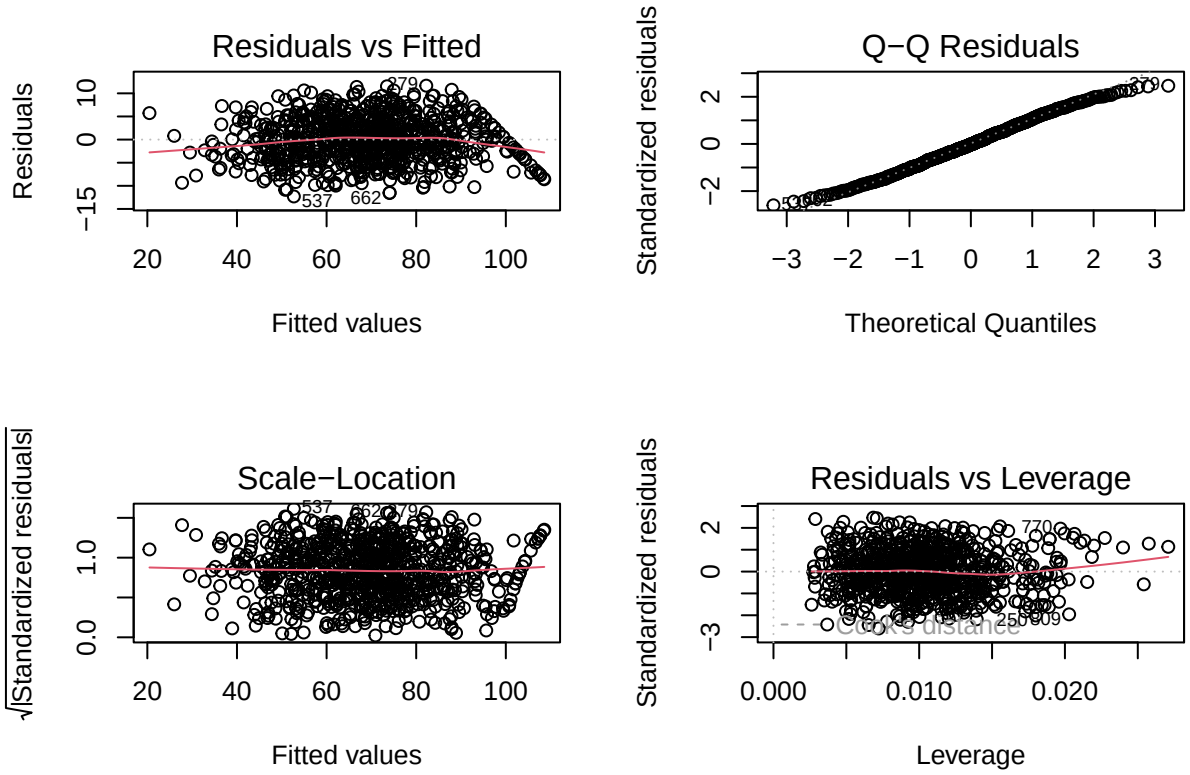
- Artık 772 değerimiz kaldı.
- Şimdi bunun için de ayrı model oluşturalım;

```r
trainsetrem_model8_rem <- lm(exam_score~. -internet_quality -gender -part_time_job -age -parental_educa
summary(trainsetrem_model8_rem)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender - part_time_job -
##     age - parental_education_level - diet_quality - extracurricular_participation,
##     data = trainsetrem_cooks_rem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2957  -3.1940  -0.1084   3.4148  11.6562
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.50388    1.95883   2.810  0.00508 **
## study_hours_per_day     9.66257    0.12114  79.763  < 2e-16 ***
## social_media_hours     -2.50919    0.15001 -16.727  < 2e-16 ***
## netflix_hours          -2.29734    0.15868 -14.478  < 2e-16 ***
## attendance_percentage   0.14403    0.01854   7.768 2.58e-14 ***
## sleep_hours             1.97576    0.14024  14.088  < 2e-16 ***
## exercise_frequency      1.49335    0.08416  17.745  < 2e-16 ***
## mental_health_rating    1.95971    0.06062  32.329  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.741 on 762 degrees of freedom
## Multiple R-squared:  0.9157, Adjusted R-squared:  0.9149
## F-statistic:  1182 on 7 and 762 DF,  p-value: < 2.2e-16
```

- Önceki modelimize göre hataların azaldığını Adjusted R kare değerinin biraz arttığını görüyoruz. Aslında modelimiz iyileşti fakat varsayım kontrollerini de yapmalıyız.

```r
par(mfrow = c(2,2))
plot(trainsetrem_model8_rem)
```



- Normallik eğrisinin uçlarda bozulduğunu görüyoruz. - Hemen testleri uygulayalım;

```
bptest(trainsetrem_model8_rem)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trainsetrem_model8_rem
## BP = 3.8199, df = 7, p-value = 0.8003
```

```
shapiro.test(residuals(trainsetrem_model8_rem))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(trainsetrem_model8_rem)
## W = 0.995, p-value = 0.01287
```

```
dwtest(trainsetrem_model8_rem)
```

```
##
##  Durbin-Watson test
##
## data:  trainsetrem_model8_rem
## DW = 1.8902, p-value = 0.06353
## alternative hypothesis: true autocorrelation is greater than 0
```

- Normalliğin artık sağlanmadığını shapiro ile görüyoruz.

10-) İki Modelin Tahminlerini Kontrol Edelim

```
tahminler_trainsetrem <- predict(trainsetrem_model8,testset)
tahminler_trainsetrem_rem <- predict(trainsetrem_model8_rem,testset)
```

- Şimdi bu iki modelin RMSE, MAE değerlerini kıyaslayalım.

```
RMSE(tahminler_trainsetrem,testset$exam_score)
```

```
## [1] 5.285712
```

```
RMSE(tahminler_trainsetrem_rem,testset$exam_score)
```

```
## [1] 5.272113
```

- Cooks ile aykırıları attığımız modelin RMSE değeri bir tık daha düşük geldi.
- Tahmin anlamında daha yakın tahminler yapan bir model olduğunu söyleyebiliriz fakat aralarında çok çok minimal bir fark var.

```
MAE(tahminler_trainsetrem,testset$exam_score)
```

```
## [1] 4.211304
```

```
MAE(tahminler_trainsetrem_rem,testset$exam_score)
```

```
## [1] 4.205315
```

- MAE değerleri de aynı şekilde çok minimal farka sahip. Küçük bir fark ile cook değerlerini attığımız, yani varsayımları sağlamayan model daha iyi tahmin yapmış.
- Yani kısaca eğer normalliği sağlasaydı cook değerlerini attığımız modeli nihai model alabilirdik fakat halihazırda cook değerlerimiz çok küçük ve modelin normalliğini bozduğundan dolayı nihai modeli mahalanobis ile temizlenmiş fakat cooks distance ile belirlenmiş outlierlar atılmamış model olarak alıyoruz. Yani trainsetrem_model8 oluyor.
- Son modelimizi incelersek;

```
summary(trainsetrem_model8)
```

```
##
## Call:
## lm(formula = exam_score ~ . - internet_quality - gender - age -
##     diet_quality - part_time_job - parental_education_level -
##     extracurricular_participation, data = trainsetrem)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -13.5681  -3.2895  -0.2003   3.6002  16.0430
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.28920    2.09467   2.525   0.0118 *
## study_hours_per_day     9.63669    0.12817  75.186  < 2e-16 ***
## social_media_hours     -2.50990    0.16146 -15.545  < 2e-16 ***
## netflix_hours          -2.21606    0.17029 -13.013  < 2e-16 ***
## attendance_percentage   0.14661    0.01999   7.334 5.57e-13 ***
## sleep_hours             1.95868    0.14916  13.132  < 2e-16 ***
## exercise_frequency      1.52780    0.09037  16.906  < 2e-16 ***
## mental_health_rating    1.95432    0.06494  30.096  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.184 on 788 degrees of freedom
## Multiple R-squared:  0.902,  Adjusted R-squared:  0.9011
## F-statistic:  1036 on 7 and 788 DF,  p-value: < 2.2e-16
```

-Bağımsız değişkenlerimiz, bağımlı değişkenimiz olan exam_score'u %90.08 oranında açıklyabiliyor yani gerçekten uygun bir model.

- Öğrencilerin günlük ders çalışma süresi 1 saat arttığında sınav puanı 9.66 puan artış gösteriyormuş yani en büyük etkenin bu olduğunu söyleyebiliriz.
- Günlük sosyal medyaya harcanan her saat sınav puanını 2.47 puan düşürüyormuş.
- Benzer bir şekilde Netflix izleme süresi de her 1 saat artışta 2.21 puan düşürüyor sınav puanımızı.
- Aslında derse katılım yüzdesinin de büyük bir etkisi var diyebiliriz. Çünkü yüzde olarak her %1 artış sınav puanını 0.14 puan artırmış.
- Günlük uyku saatindeki 1 saatlik artış sınav puanına +1.96 puan olarak etki etmiş.
- Egzersiz sıklığı 1 puan arttıkça sınav puanı yaklaşık +1.52 puan etkilenmiş.
- Akıl sağlığı ise 1 birim arttığında sınav puanı +1.96 puan etkilenmiş.