Econometrics I - Case Study 1

Group 3, Instructor: Univ.Prof. David Preinerstorfer, Ph.D

Julian Hofmaninger, Tsz Lam Hung and Daniel Diederichs

October 20, 2025

Problem 1: Data aquisition

Loading data

```r
library(readxl)
data <- read_xlsx("CA_Schools_EE14/CASchools_EE141_InSample.xlsx")
CAschool <- subset(data, charter_s == 0)
```

Describing variables

- `testscore`: Measures the sum of the scores on exams in Maths and English of 5th grade students from a school.
- `str_s`: Describes how many students come per teacher in a school.
- `med_income_z`: Represents the median income considering all inhabitants of the schools district that are at least 15 years old.

Predicted correlation

- med_income_z: We would argue that a higher `med_income_z` leads to higher values of `testscore` because in wealthier neighbourhoods people tend to be better educated and therefore can provide their children with better assistance in school related matters.

- str_s: We believe that fewer students per teacher have a positive effect on the value of `testscore` as fewer pupils per teacher means that the teachers can put more emphasis on the needs of the fewer students they have and might be less in a hurry of moving from one class to another to take time and actually go through topics that were not fully or extensively enough covered. Therefore, `str_s`would be negatively correlated to the value of `testscore`.

Computed correlation

```r
cor(CAschool$testscore, CAschool$med_income_z) # 0.5950106
```

```
## [1] 0.5950106
```

The computed empirical correlation between `testscore` and `med_income_z` is a positive number. This suggests that a higher median income leads to better test results of students and therefore supports the argument we made.

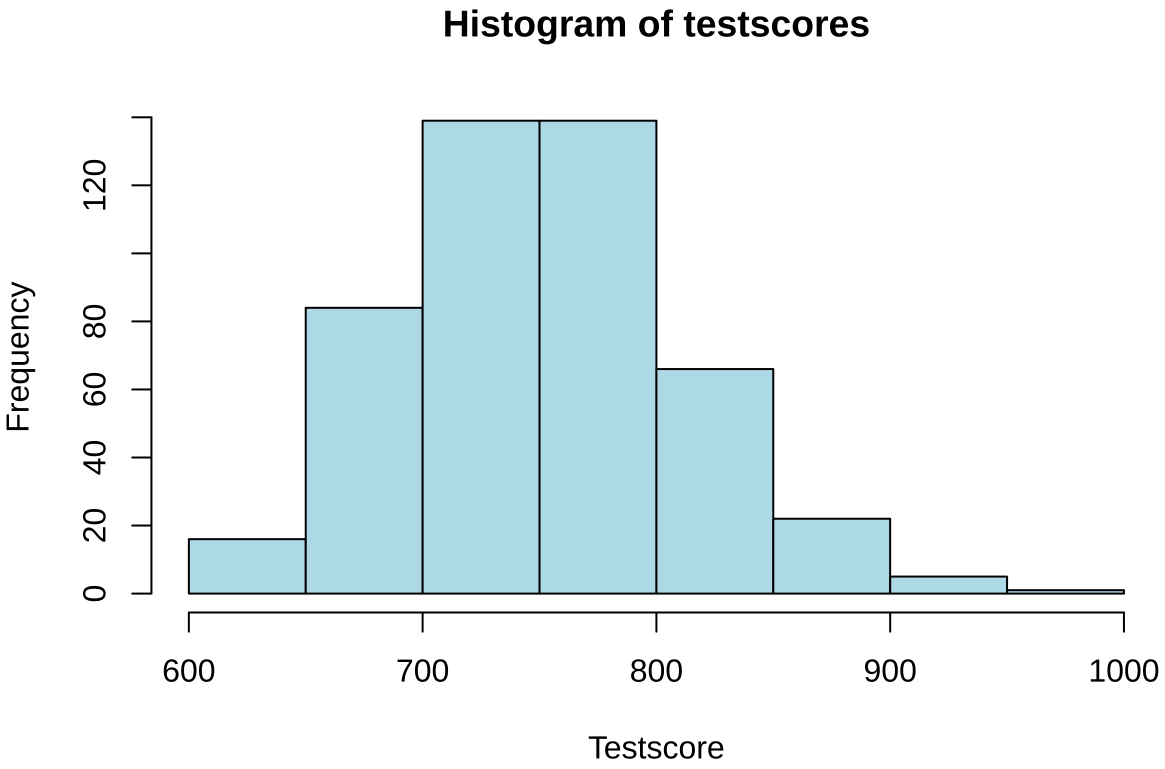
```
cor(CAschool$testscore, CAschool$str_s) # -0.05576373
```

```
## [1] -0.05576373
```

The computed empirical correlation between `testscore` and `str_s` is a negative number. This supports the argument of suggesting that fewer students per teacher, meaning a lower value for `str_s` has a positive impact on the value of `testscore`.

Problem 2: Descriptive statistics

Histogram of testscores



Analyse extreme values

```
minTestscore <- CAschool[which.min(CAschool$testscore), ] # Westmorland Elementary
minTestscore$schoolname
```

```
## [1] "Westmorland Elementary"
```

```
maxTestscore <- CAschool[which.max(CAschool$testscore), ] # Tom Matsumoto Elementary
maxTestscore$schoolname
```

```
## [1] "Tom Matsumoto Elementary"
```

```
print(paste(minTestscore$schoolname, ": ",
            minTestscore$te_salary_avg_d, " | ", maxTestscore$schoolname,
            maxTestscore$te_salary_avg_d)) # Min: 59715 | Max: 80971
```

```
## [1] "Westmorland Elementary :  59715  |  Tom Matsumoto Elementary 80971"
```

- `te_salary_avg_d`: The average salary of a teacher at at school in the district of the school with the lowest value for `testscore` is by far lower than the average salary of a teacher in the district of the school with the maximum value for `testscore`. This indicates that at the school with the worst values for `testscore` teachers might be worse or at least not as motivated since their salary is on average lower.

```
cat(paste0(minTestscore$schoolname, ": ", minTestscore$str_s,
           " |\n", maxTestscore$schoolname, ": ",
           maxTestscore$str_s)) # Min: 15.82609 | Max: 25.35294
```

```
## Westmorland Elementary: 15.8260869979858 |
## Tom Matsumoto Elementary: 25.3529415130615
```

- `str_s`: The student teacher ratio is lower at the school with the worst value for `testscore` than at the school with the highest value for `testscore`. That means that at the school that scores the lowest on test there are fewer students per teacher. That is actually quite interesting as we have spotted a negative correlation before meaning that values for `testscore` tend to be higher with a lower value for `str_s`

```
print(paste(minTestscore$schoolname, ": ",
            minTestscore$med_income_z, " | ", maxTestscore$schoolname, ": ",
            maxTestscore$med_income_z)) # Min: 15000 | Max: 51556
```

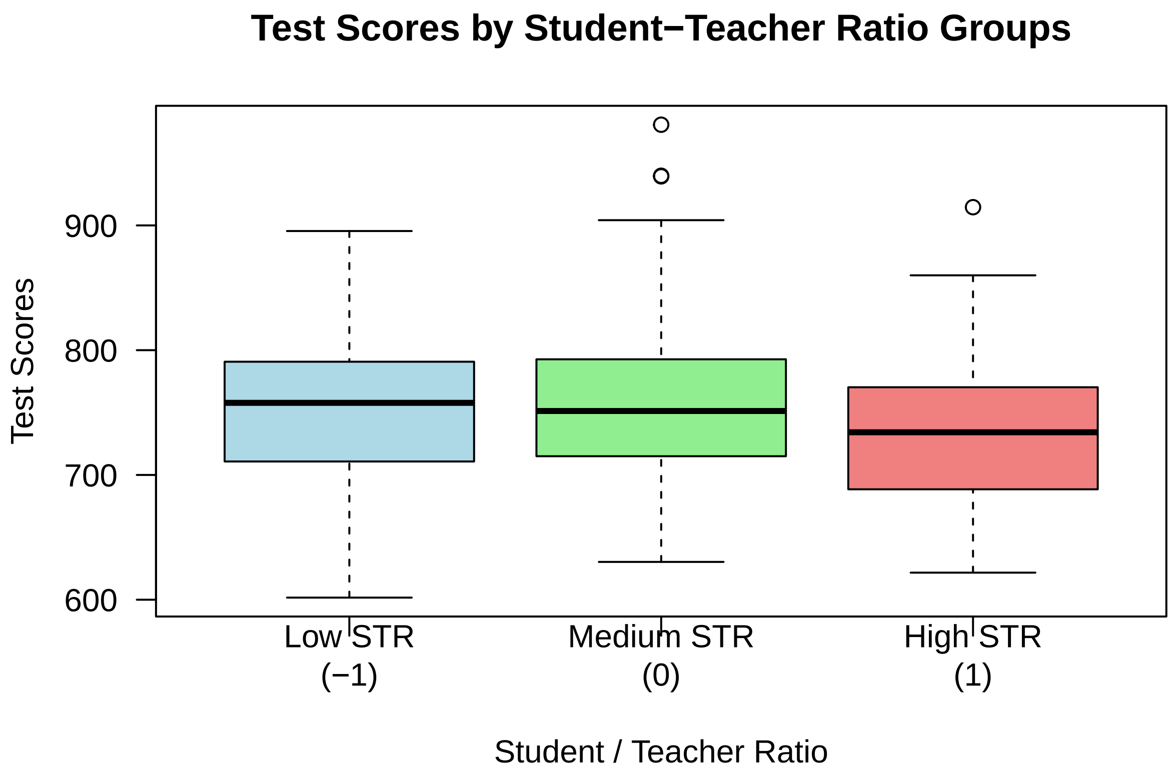
```
## [1] "Westmorland Elementary :  15000  |  Tom Matsumoto Elementary :  51556"
```

- `med_income_z`: The median income in the district of the school that has the students that score the lowest on exams is by far lower than the median income in the district of the school that has the students with the highest scores on exams. As already noticed before when computing the empirical correlations this confirms that there is a tendency of higher value on `testscore` in wealthier areas.

Problem 3: Plots

```
q20 <- quantile(CAschool$str_s, 0.2)
q80 <- quantile(CAschool$str_s, 0.8)
CAschool$group_strs <- ifelse(CAschool$str_s < q20, -1,
                        ifelse(CAschool$str_s > q80, 1, 0))
```

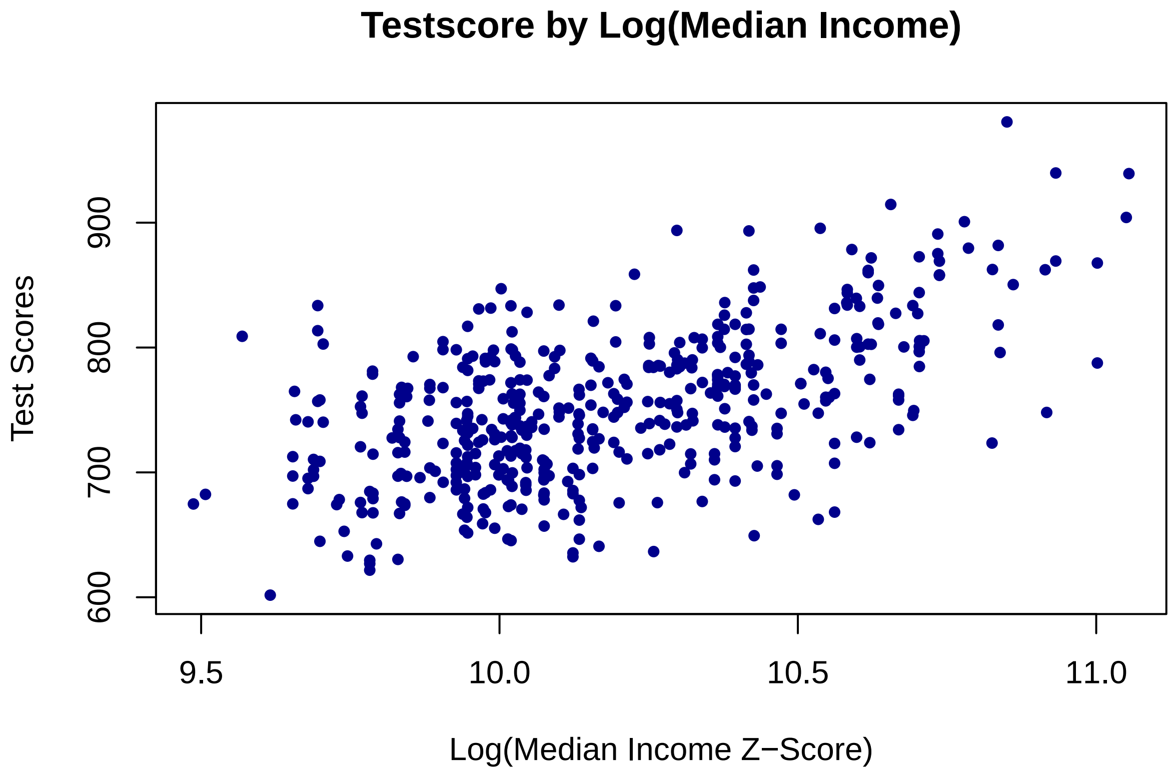
```
boxplot(testscore ~ group_strs, data = CAschool,
        names = c("Low STR\n(-1)", "Medium STR\n(0)", "High STR\n(1)"),
        main = "Test Scores by Student-Teacher Ratio Groups",
        ylab = "Test Scores",
        xlab = "Student / Teacher Ratio",
        col = c("lightblue", "lightgreen", "lightcoral"),
        las = 1)
```



The boxplots show that the value for `tescore` tends to be lower with a higher student per teacher ratio. Even though it is interesting to observe that the absolutely highest scores of schools are achieved by schools that belong to the group with a medium or high student to teacher ratio.

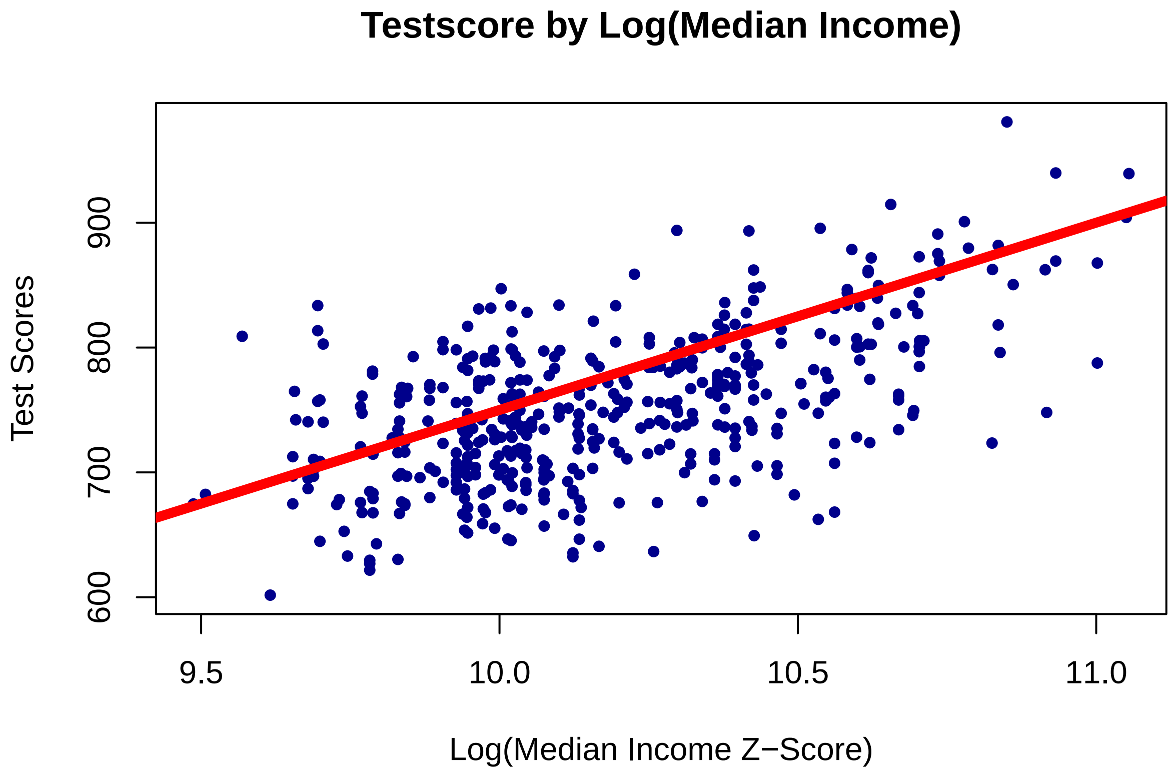
Problem 4: Scatterplot

```
CAschool$lmed_income <- log(CAschool$med_income_z)
plot(CAschool$lmed_income,CAschool$testscore,type="p",
     main="Testscore by Log(Median Income)",
     ylab="Test Scores",
     xlab="Log(Median Income Z-Score)",
     pch=16,
     col="darkblue",
     cex=0.8)
```



From the scatterplot with `testscore` on the y-axis the natural logarithm of the median income on the x-axis gives a good illustration of a potential linear relationship between these two variables.

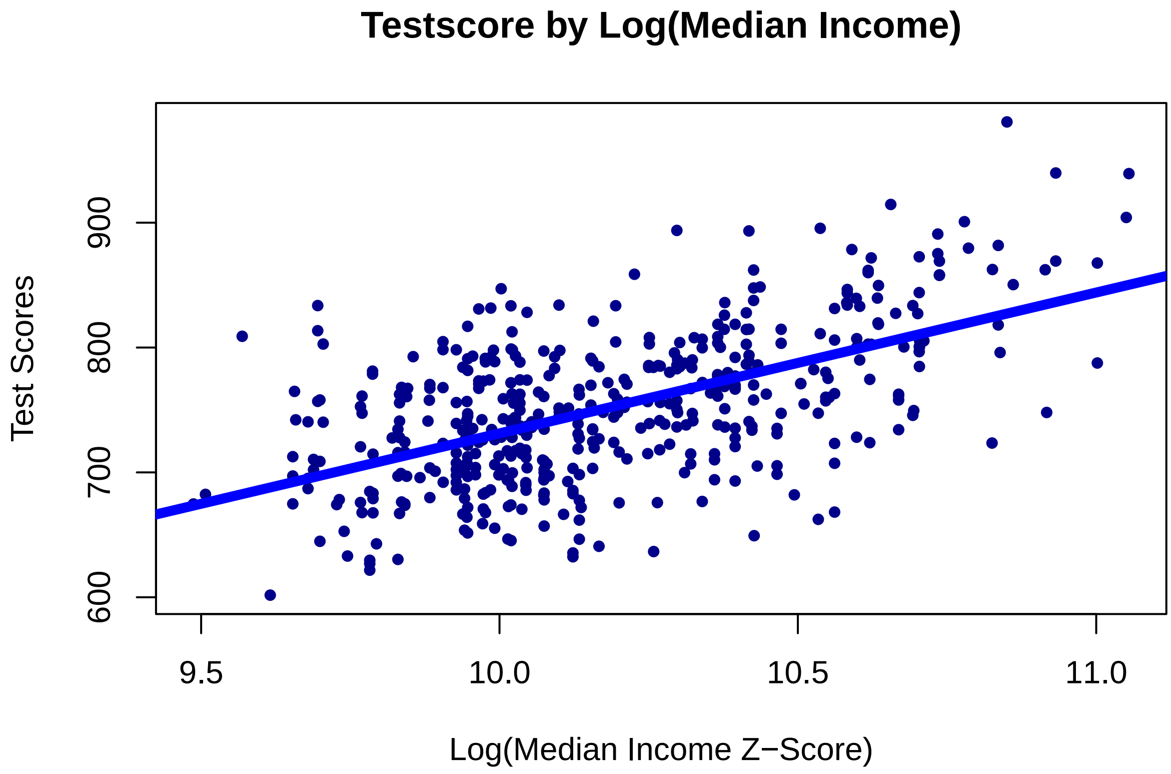
```r
plot(CAschool$lmed_income,CAschool$testscore,type="p",
     main="Testscore by Log(Median Income)",
     ylab="Test Scores",
     xlab="Log(Median Income Z-Score)",
     pch=16,
     col="darkblue",
     cex=0.8)
abline(a=-750, b=150, col="red",lwd=5)
```



For the estimation of the regression line we took the points $x = 10$, $y = 750$ and $x = 11$, $y = 900$. Therfore, we get a regression line of $y \sim 150 * x - 750$.

Problem 5: OLS estimation and illustration

```
plot(CAschool$lmed_income,CAschool$testscore,type="p",
     main="Testscore by Log(Median Income)",
     ylab="Test Scores",
     xlab="Log(Median Income Z-Score)",
     pch=16,
     col="darkblue",
     cex=0.8)
reg1 <- lm(testscore ~ log(med_income_z), data=CAschool)
abline(reg1,col="blue",lwd=5)
```



Read from chart

Reading from the chart we get a value of approximately 725.

Calculating by hand

```
112.8 * log(22026.47) - 396.8
```

```
## [1] 731.2
```

Calculating by hand we have to insert the value in the regression formula from above which gives: $f(22026.47) = 112.8 * log(22026.47) - 396.8$ and that further results in: $f(22026.47) = 731.2$

Using the predict function

```
predict(reg1, newdata=data.frame(med_income_z=c(22026.47)))
```

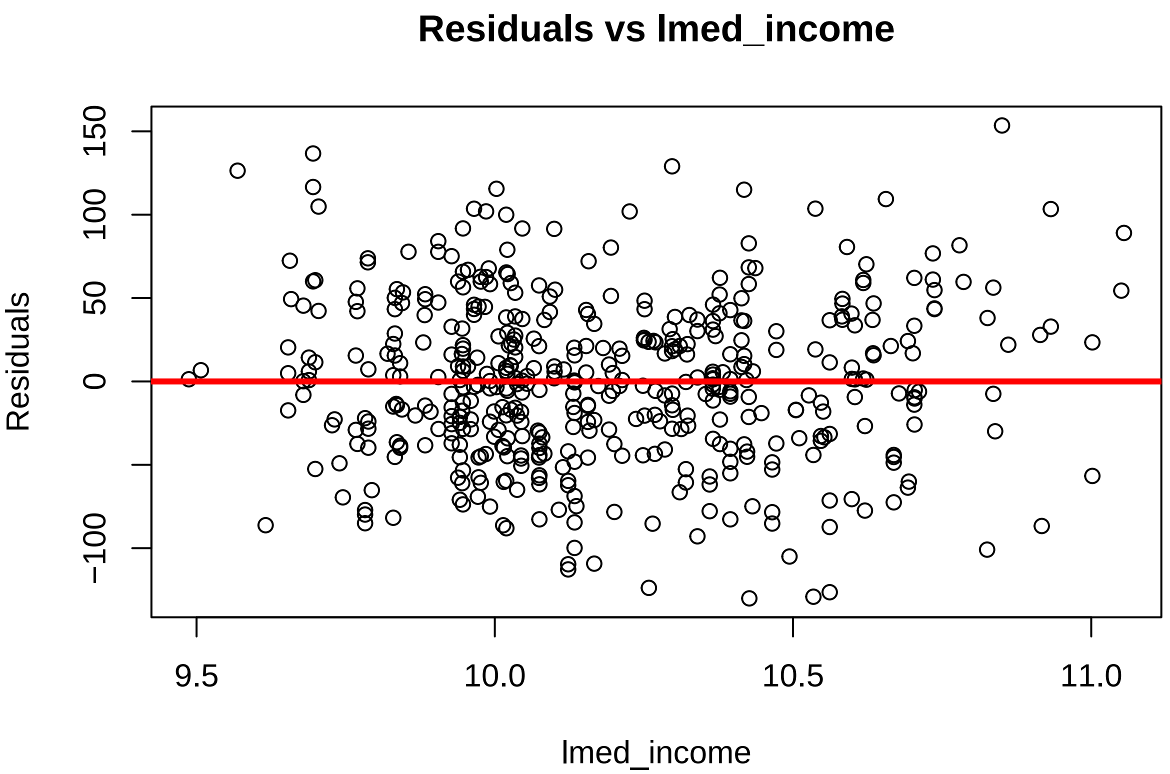
```
##        1
## 731.2601
```

```
predict(reg1, newdata=data.frame(med_income_z=c(22026.47*exp(-0.5))))
```

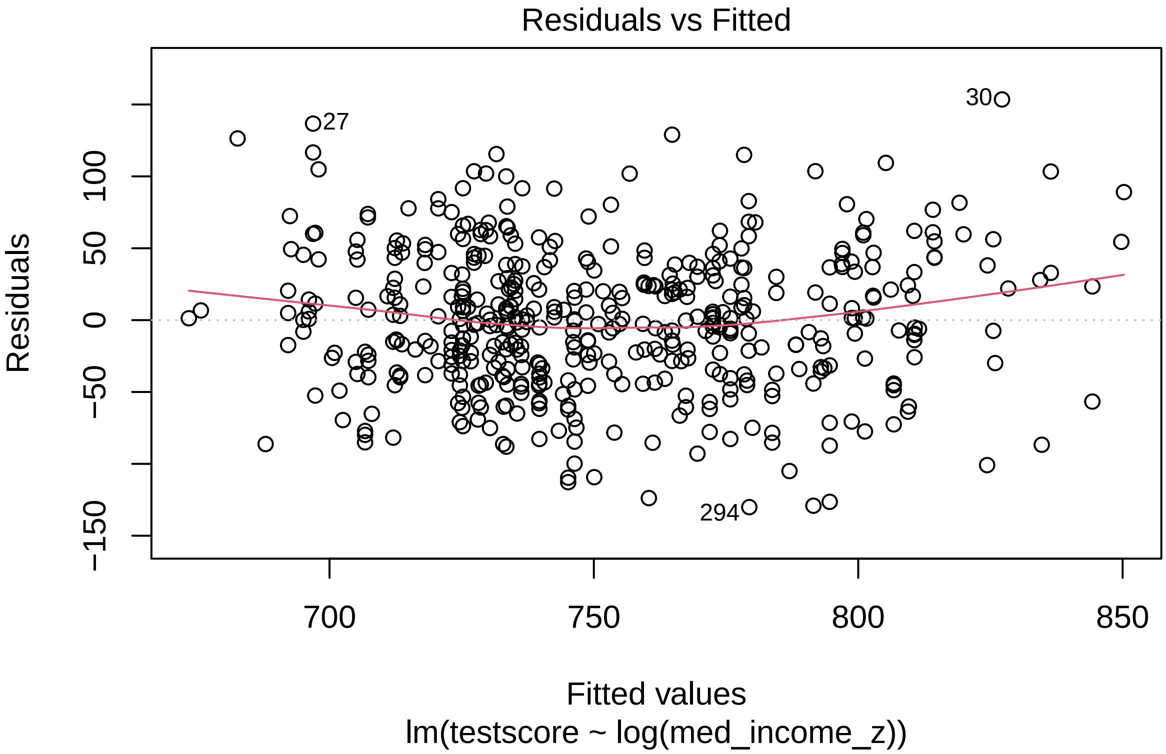
```
##        1
## 674.8564
```

Problem 6: Residual plot

```
residuals <- resid(reg1)
plot(CAschool$lmed_income, residuals,
     xlab = "lmed_income",
     ylab = "Residuals",
     main = "Residuals vs lmed_income")
abline(h=0, col="red", lwd=3)
```



```
plot(reg1, which = 1)
```



The spread of the residuals plotted against the regressor `lmed_income_z` remains rather constant across different levels of the regressor as well as different levels for the fitted value as we can see in the second chart. It is not possible to clearly identify a cone, fan or balloon shape and therefore we would suggest that the homoskedasticity assumption holds. However, what we can observe in the second chart that the red line (LOESS-curve) shows a slight curve. This suggests that the relationshop between the dependent and independent variable is not fully linear. To mitigate this effect we could transform the dependent variable and apply `log` as well to define a `log-log` model like: $log(testscore) \sim log(med\_income\_z)$.