**STATISTICS OF UDEMY**

**COURSES**

A FINAL PROJECT REPORT SUBMITTED

IN FULFILLMENT OF THE REQUIREMENTS FOR COURSE

STAT 250 – APPLIED STATISTICS

DEPARTMENT OF STATISTICS OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

Mustafa ÇANDIR 2290583

July 2022

**ABSTRACT**

In today's world, everyone wants to be their own boss. At this point, people are starting to educate themselves. How are they doing it? Platforms like udemy become important. People are using or paying for courses to learn new skills on the platform. These platforms include different levels of the learner. You can find hundreds of lectures on different subjects. Some factors affect the selling of courses. If the content duration is long, it probably will not sell many. Also, the number of viewers decreased for these courses. In the light of these thoughts, we analyze Udemy Course data to see how factors are affected by other variables. We perform some tests and analyses of the data. Exploratory data is used.

1. INTRODUCTION

1.1. Data description

1.2. Research questions

1.3. Aim of the study

2. METHODOLOGY

3. RESULTS AND FINDINGS

3.1 Hypothesis test

3.2 One Sample t-test

3.3 Two-sample t-test

3.4 One-way Anova

3.5 Simple Linear Regression

3.6 Multiple Linear Regression

4. DISCUSSION/CONCLUSION

5. REFERENCES

# 

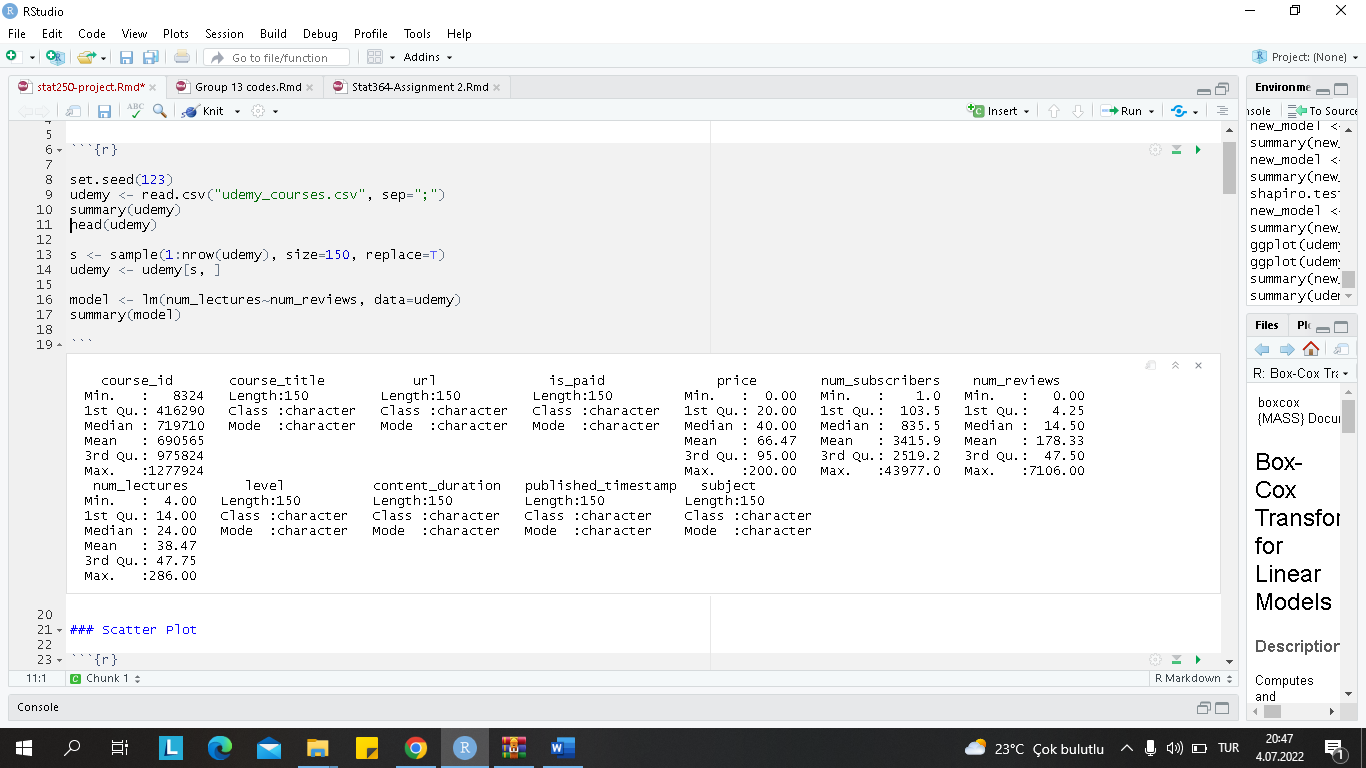
# 

# **INTRODUCTION**

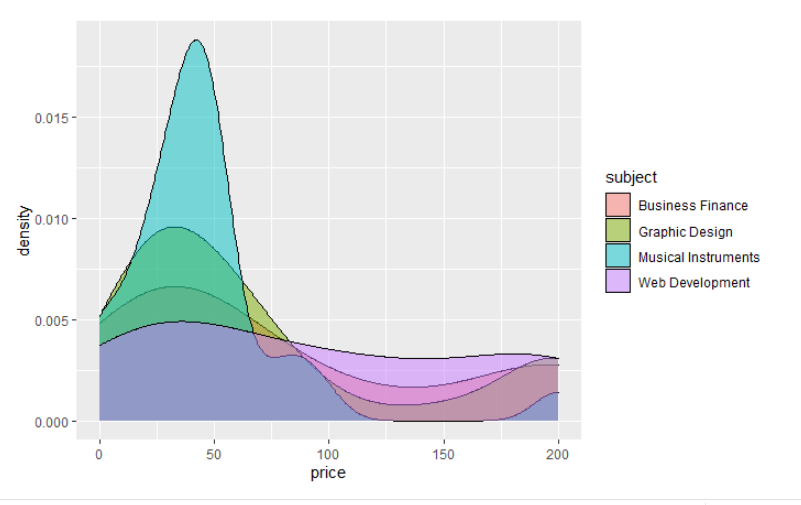
This study aims to show the relations between variables in udemy courses data. In this study, hypothesis tests, simple and multiple linear regression, ANOVA were utilized to examine relationships between variables. In this analysis, we tested whether the average price paid for courses is different from 40 and whether average number of subscribers and the number of reviews are equal to each other or not. Also, we try to see whether levels have an impact on course price or not. In addition, with multiple linear regression, we try to find out whether some factors have a relationship with the number of lectures or not. Besides, simple linear regression examines whether there is a relationship between the number of courses and the number of reviews of Udemy courses.

## **1.1. Data description**

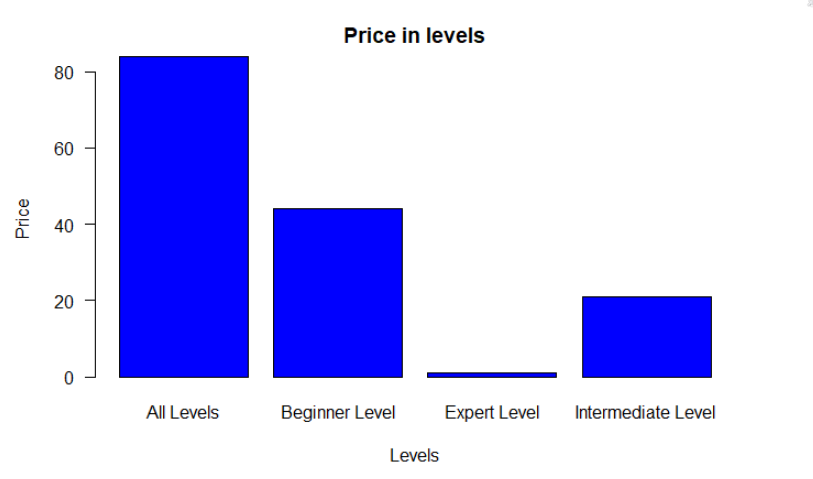
This dataset contains 3.678 observations taken from Udemy. Udemy is an online course platform where anybody can create or buy a course. The columns' names are; "course\_id", "course\_title", "url", "is\_paid", "price", "num\_subscribers", "num\_reviews", "num\_lectures, "level", "content\_duration", "published\_timestamp", "subject". The data contains 7 characters and 5 integers. The summary of data is:



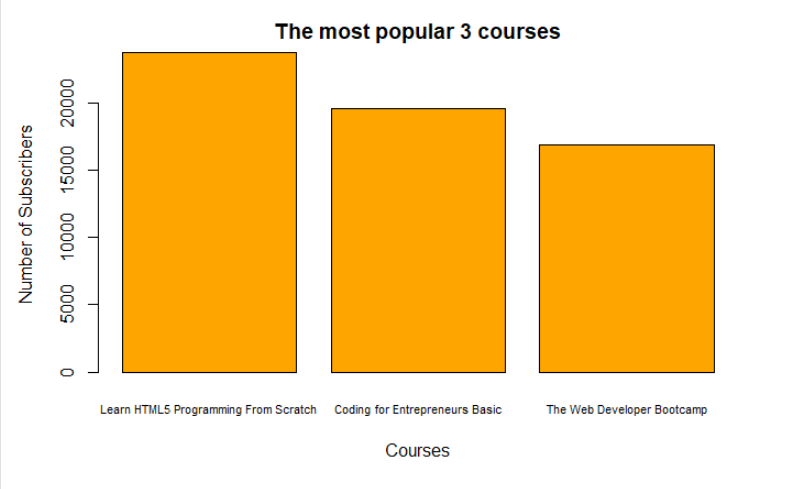
*Figure 1. Data Summary*



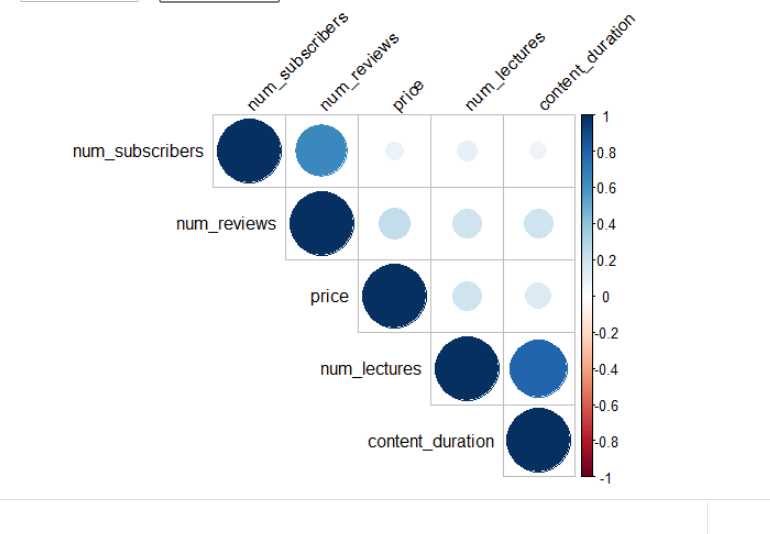
*Figure 2. Price Density by Subject*



*Figure 3. Price in Levels Table*



*Figure 4. The Most Popular Courses Table*



*Figure 5. Correlation Matrix*

## **1.2. Research questions**

1.2.1 Is there a relationship between the price of courses and the number of lectures on udemy courses?

1.2.2 Is there a relationship between the number of lectures and the number of reviews for udemy courses?

1.2.3 Is the average price paid for courses different from 40?

1.2.4 Are the average number of subscribers and the number of reviews equal to each other?

1.2.5. Is there any difference between the average price for different levels?

1.2.6. Does the number of subscribers, course price, number of reviews, content duration and does money paid for courses or not affect the number of lectures?

## **1.3. Aim of the study**

Study aims to test whether the relationship between the number of lectures and the number of reviews for udemy courses or not. In addition, it examines any difference between the average prices.It is also examined how many factors such as the relationship between the average number of subscribers and the number of comments, course prices, and the number of subscribers affect the number of course content. In other words, the general purpose of this study is to examine the factors affecting udemy courses.

# **2. METHODOLOGY**

In this project, RStudio is used to create plots and analyses. In the analysis parts that include hypothesis test, simple linear regression, multiple linear regression, ANOVA, some functions found in R such as aov(), t.test(), and lm() are used.

# **3. RESULTS AND FINDINGS**

## **3.1 Hypothesis test for Price of Lectures and the Number of Lectures**

**Q.** Is there a relationship between the price of courses and the number of lectures on udemy courses?

**Response:** The price and number of lectures

**Proposed method:** z-Test (Hypothesis Test)

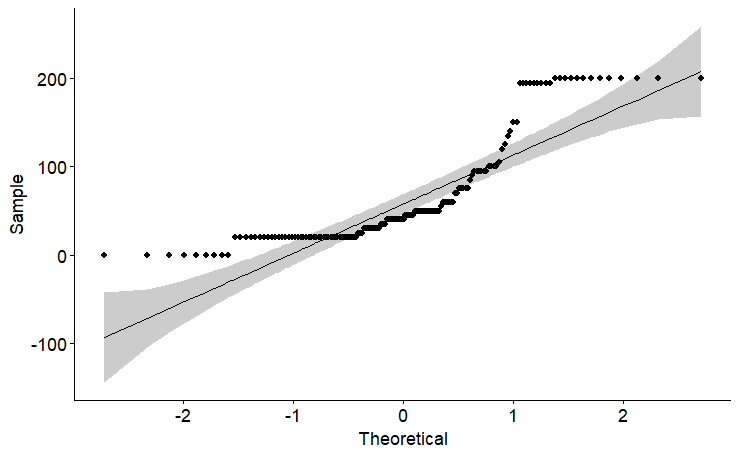
In order to test whether the assumption is true or not, some examinations were made on the data first.

And then, create a hypothesis test,

H0: μ1=μ2 (The average of price and the number of lectures)

H1: μ1≠μ2 (The average of the price and the number of lectures)

### **Assumption Check:**



*Figure 6. QQ Plot for Price*

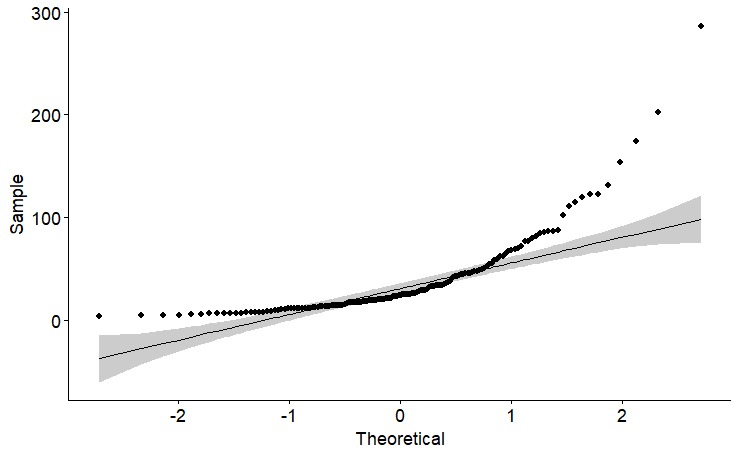
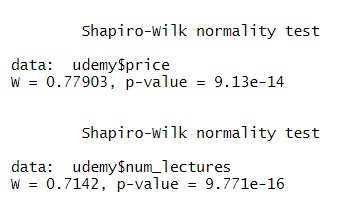
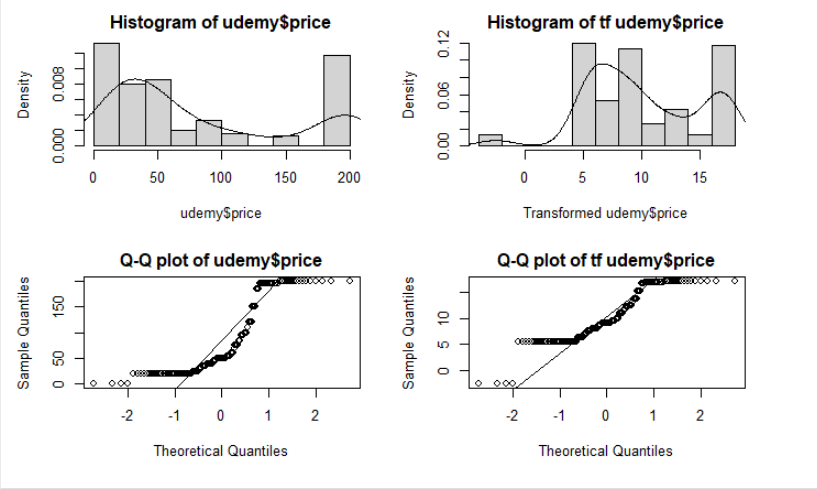


Figure 2. *QQ Plot for Number of Lectures*



After normality and Shapiro Wilk tests, we see that observations are not distributed normally. So, we apply box cox transformation.



#### Figure 3. Box Cox Transformation for Price

#### es

Figure 4. Box Cox Transformation for Number of Lectures

We tried transformation methods for positively skewed data. However, none of them were distributed normally. We cannot apply the parametric methods so we should try the non-parametric methods.

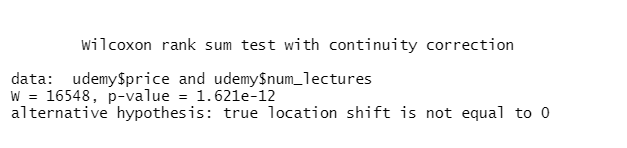


Figure 5. Wilcoxon test

The p-value = 1.621e-12 is less than the alpha value of 0.05, so we should reject the null hypothesis . We can conclude distributions for these two are different.Thus, the means of number of lectures and price paid for courses are different from each other.

3.2 One Sample t-test

Q. Is the average price paid for courses different from 40?

**Response:** The average price paid for courses

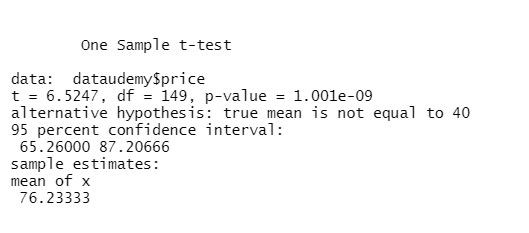
**Proposed method:** One-sample hypothesis testing

We want to test if the average price paid for courses is different from 40. To see that One sample t-test is performed.

Hypotheses are;

H0 : The average price paid for courses is 40.

H1 : The average price paid for courses is not 40.



*Figure 7. One Sample T-Test*

Since the p-value (1.001e-09) is less than 0.05, We reject the null hypothesis. So, we have enough evidence to say that the mean of course price is different than 40.

3.3 Two-sample t-test

Q. Are the average number of subscribers and the number of reviews equal to each other?

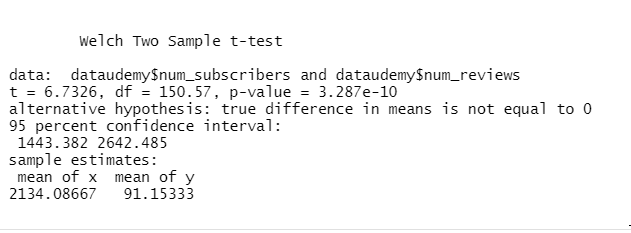
**Response:** The number of subscribers and reviews

**Proposed method:** Two-sample hypothesis testing

Hypotheses are;

H0 : The average number of subscribers and the number of reviews are equal to each other.

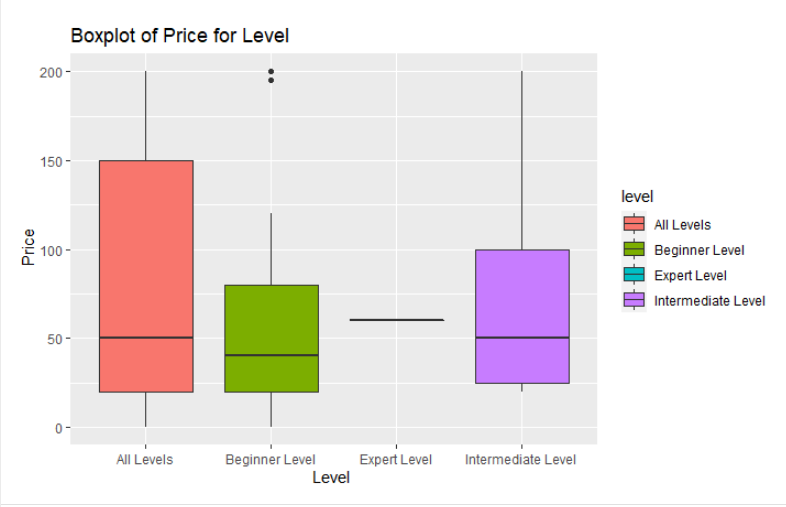
H1 : The average number of subscribers and the number of reviews are not equal to each other.



*Figure 8. Two Sample T-Test*

According to the two sample t-tests, the p (3.287e-10) value is highly smaller than 0.05 We reject the null hypothesis. So, we have enough evidence to say that the mean between the number of subscribers and the number of reviews are not equal to each other.

3.4 One-way Anova



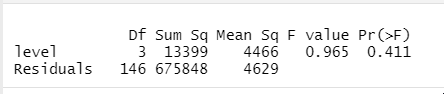
*Figure 9. One Way Anova*

The means of levels are close to each other. beginner

Q. Is there any difference between the average price for different levels?

**Response:**  Price paid for courses

**Proposed method:** ONE-WAY ANOVA



*Figure 10. One Way Anova Summary*

As a result of the test, we don't have enough evidence to reject the null hypothesis since p value (0.411) is bigger than 0.05. Thus, we can conclude that levels have no real impact on the price paid for courses.

### **Assumption Check:**

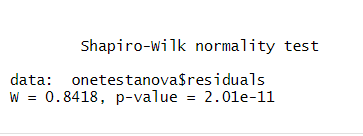
#### **1. Normality of The Residuals**

Normality for price paid for courses will be checked.

Hypotheses are;

H0: Errors are following a normal distribution.

H1: Errors are not following a normal distribution.



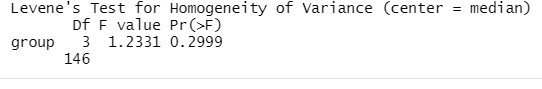
*Figure 11. Shapiro Wilk Test*

Shapiro-Wilk normality test demonstrates p value is smaller than 0.05. Thus we have enough evidence to reject the null hypothesis. So, errors are not normally distributed.

Hypotheses are;

H0: The variance between the groups is equal.

H1: The variance between the groups is *not* equal.



*Figure 12. Levene’s Test*

P value is bigger than the probability value (0.5). That's why we don't have enough evidence to reject the null hypothesis. Thus, the variance between the groups is not equal.

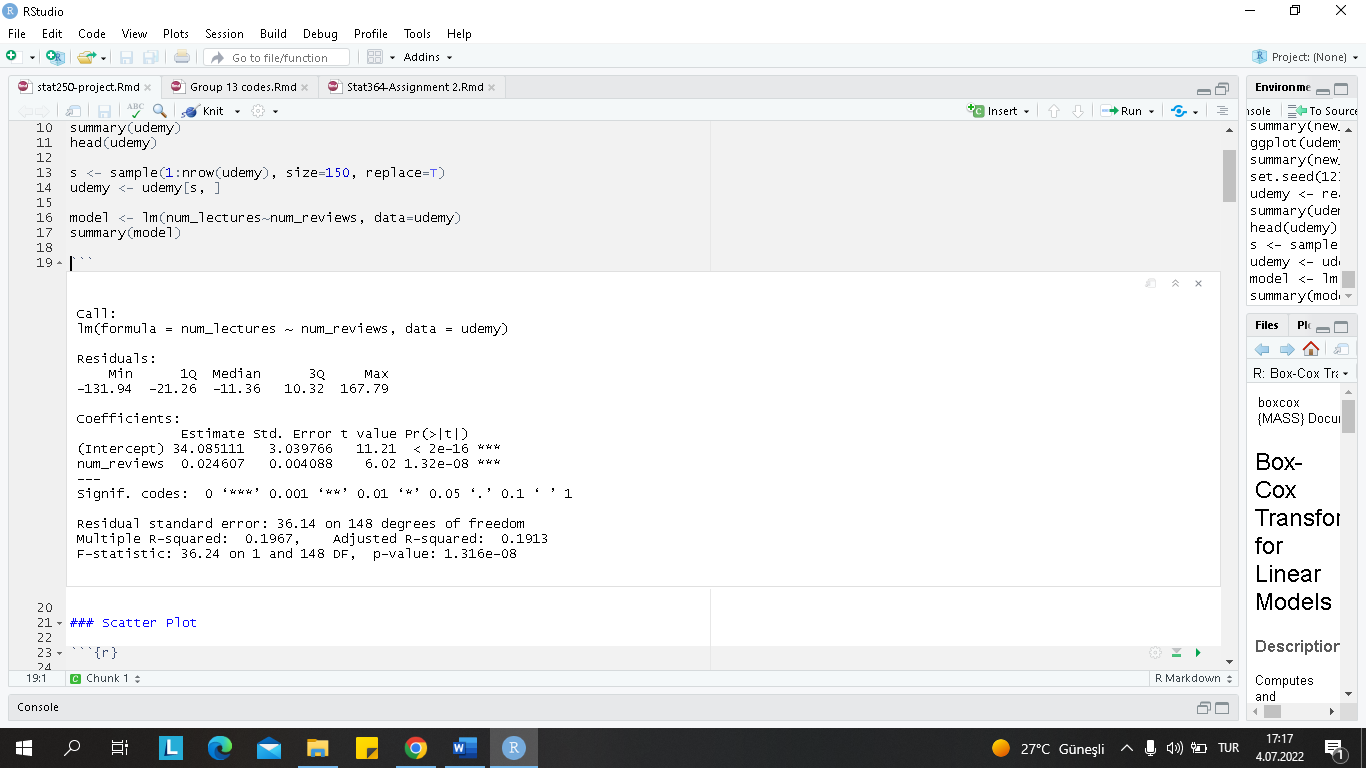
As a result of the normality and variance tests, assumptions are violated. So, we conclude that anova results are worthless. we should use different analyses.

## **3.5 Simple Linear Regression Between Number of Lectures and Number of Reviews**

**Q.** Is there a relationship between the number of lectures and the number of reviews for udemy courses?

Before starting the analysis, we can hardly guess what kind of relationship there might be between the two. We may have a weak relationship; however, we can predict that it will turn out to be positive.

In order to test whether the assumption is true, some examinations were made on the data first. A simple linear regression has been established between the X and Y variables and the summary output of the model is given below. The summary of the model gives us information about some values like residuals, estimates, standard errors, t values, and p-value values. The p-value of the number of reviews is 1.32e-08, it is less than 0.05. It means that this variable has a significant effect on the number of lectures on udemy.

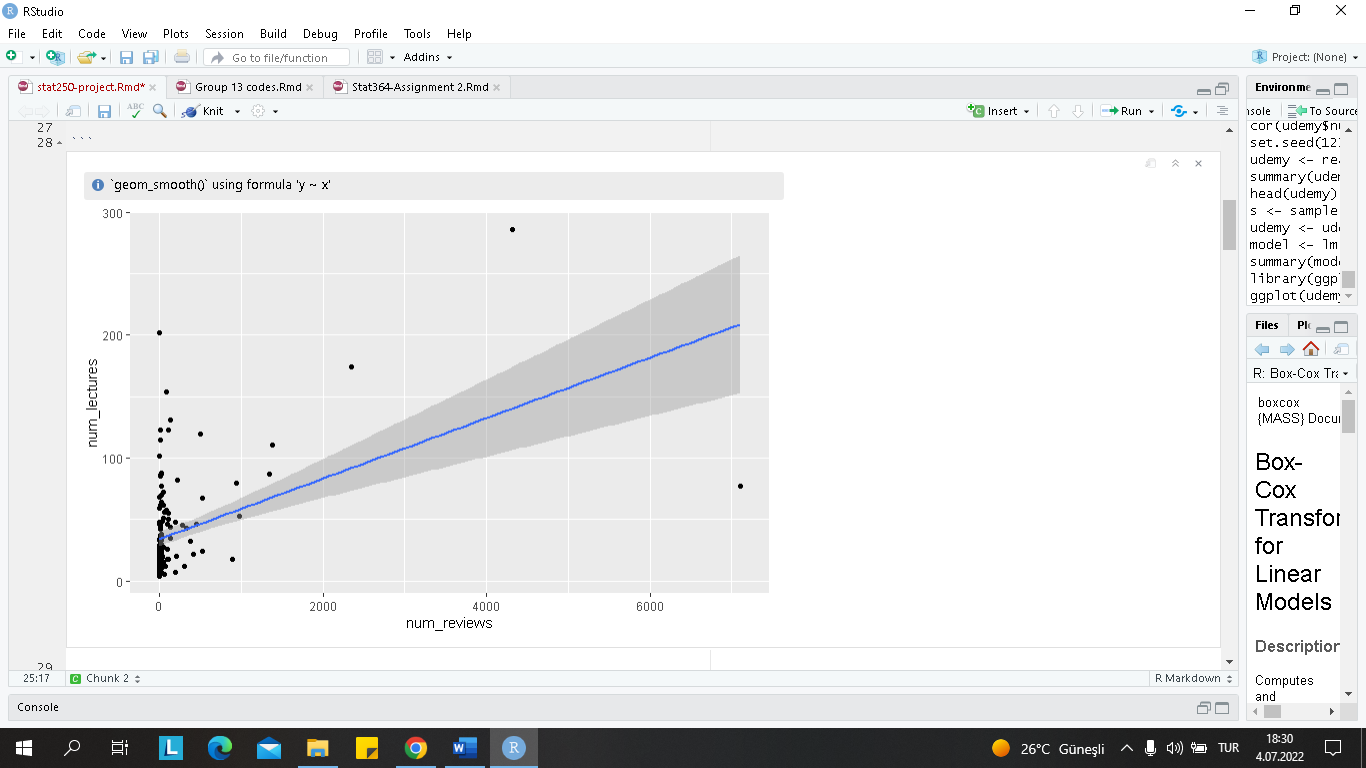


*Figure 13. Simple Linear Regression*

The 0 value is equal to 34.085111 while the 1 value is equal to 0.024607. So, if x (number of reviews) increases by 1 unit, the expected change in y is about 0.025. The effect of the number of reviews on the number of lectures is small. Besides, the r-squared adjusted value is 0.1913, which is very small too. R-square adjusted measures the linear relationship between x and y. R square adjusted value implies that most of the variability of y is not explained by the model very well. In addition, the p-value of the model is 1.316e-08, so it is smaller than 5%. This indicates that the model is significant.

### **Assumption Check:**

#### **1.** **Scatter Plot**

****

|  |
| --- |
| *Figure 14. Scatter Plot for Linear Regression* |

Linear regression requires the relationship between the response variable and independent variable to be linear. The linear relationship between these two variables seems to be positive and weak. As the number of reviews increases, the number of lectures will increase. The model has heteroskedasticity problems. Because of this reason, we need to transform the X variable.

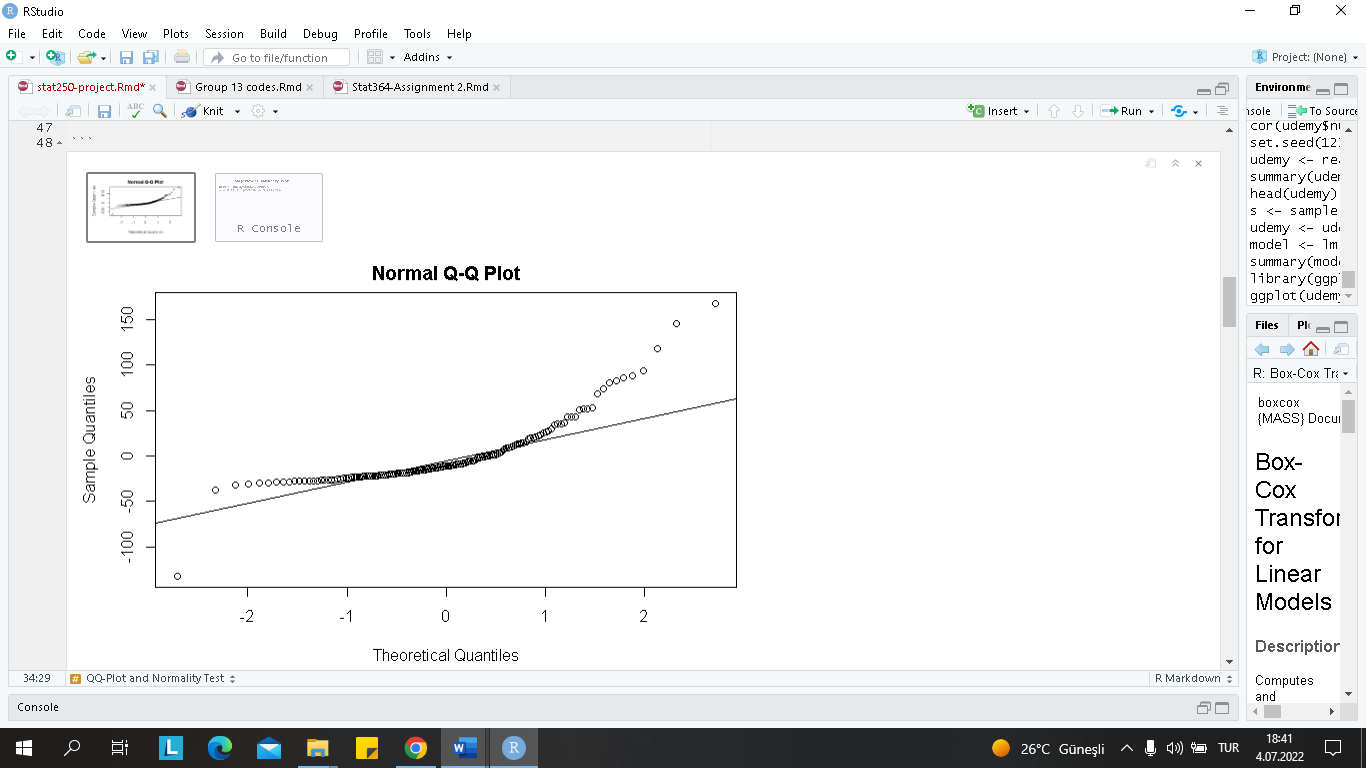
#### **2.** **QQ-Plot & Shapiro Wilk Test**

|  |
| --- |
| Here, it is checked whether errors are normally distributed with a mean zero. It uses residuals to check the assumptions for errors. The p-value of the Shapiro Wilk test is smaller than 5%, we can reject the null hypothesis, and we can say that the response variable is not normally distributed.    *Figure 15. Shapiro-Wilk Test* |

For Shapiro-Wilk Test:

H0: the response is normally distributed

H1: the response is not normally distributed

****

*Figure 16. Normal Q-Q Plot*

In addition, normality of data was checked in the QQ plot which demonstrates that residuals are not following normal distribution. Also, we can say that there are some outliers. As a result, the plot supports the result of the shapiro wilk test.

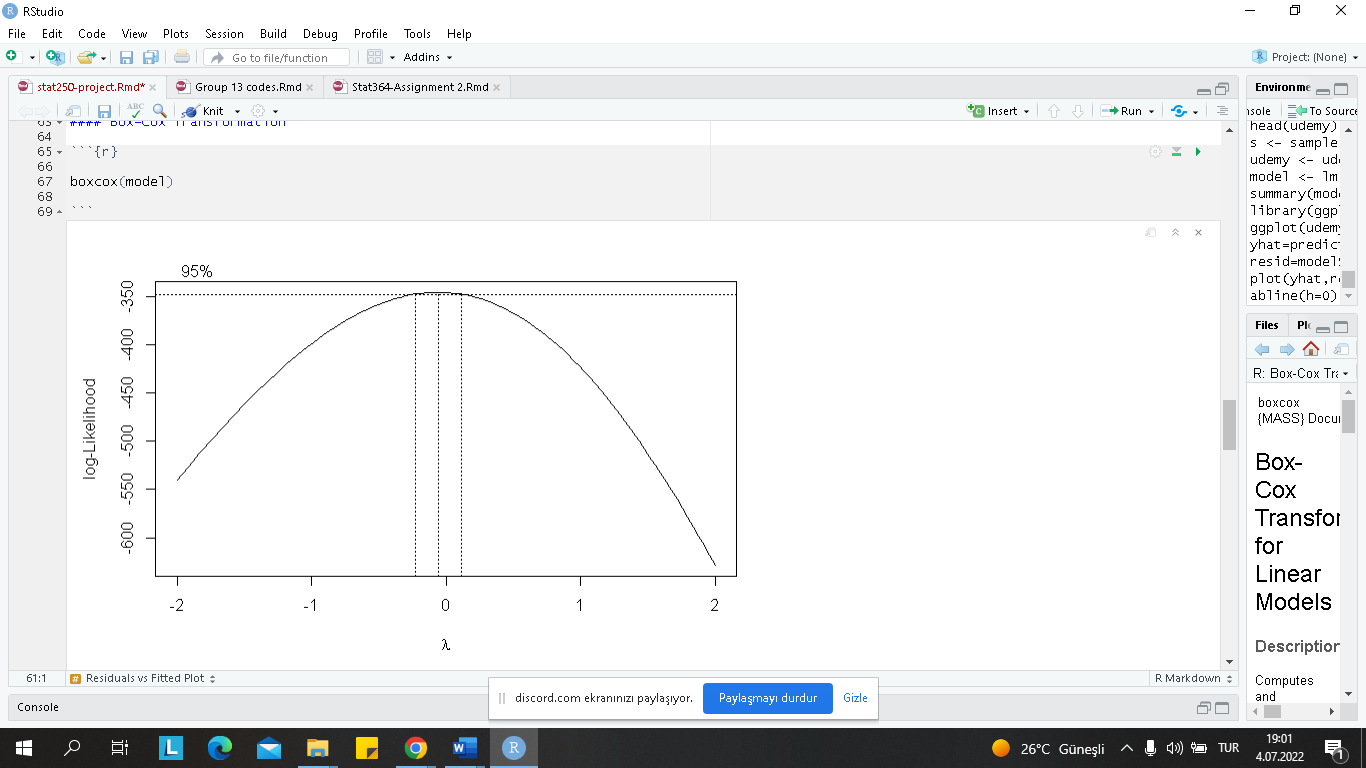
#### **3.** **Residuals and Fitted Plot**

|  |
| --- |
| *Figure 17. Residuals vs Fitted Plot* |

#### 

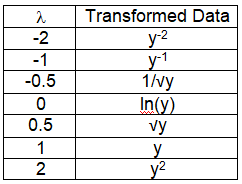
The plot of residuals versus fitted values demonstrates that the residual variance decreases as the fitted values increase. It violates the assumption of constant error variance. Also, there is a pattern here, we can consider transformation on Y.

### **Transformations**

****

*Figure 18. Box Cox Plot*

To find out which Y transformation we should do, we first need to examine the box cox plot. In the box cox plot, we can see the 95% confidence interval for λ includes 0.

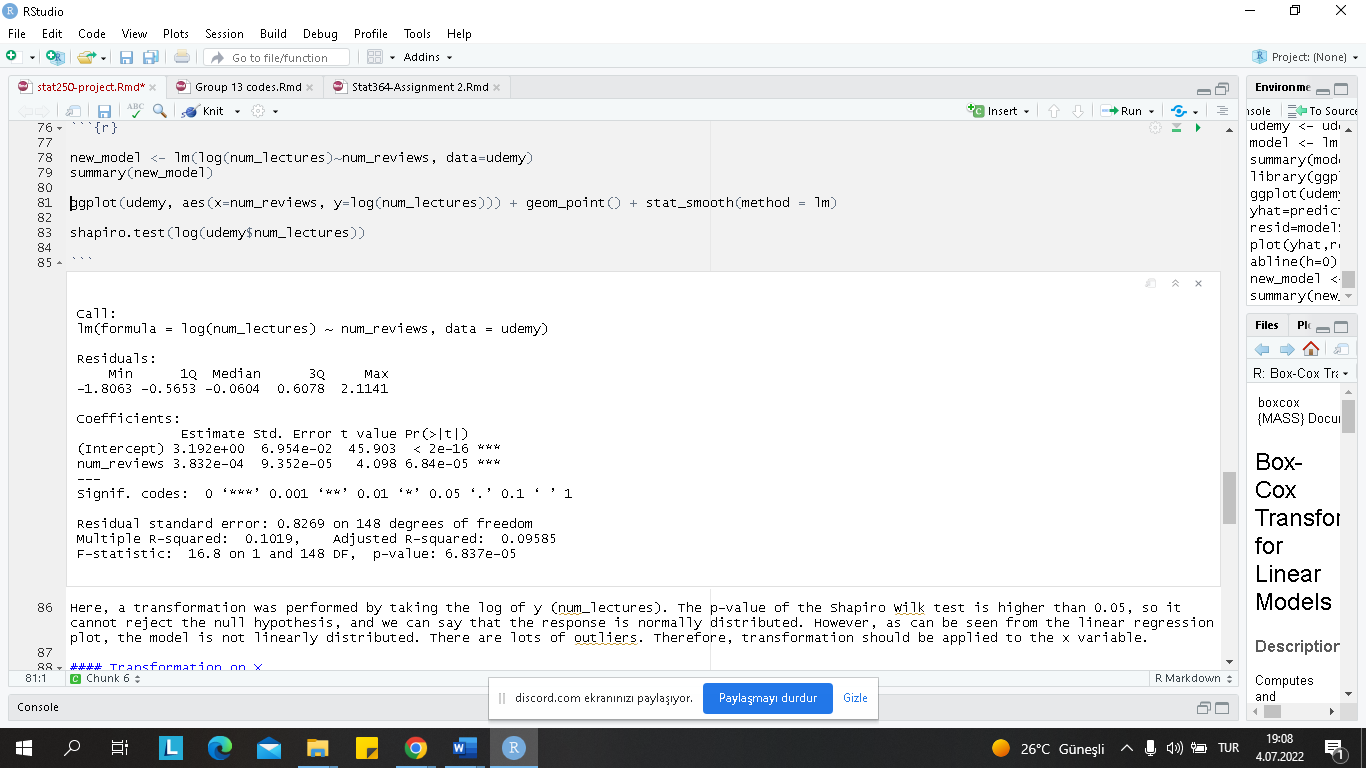


As we can see from the table on the left, if interval includes 0, then we can apply log (y) transformation.

*Figure 19: Box Cox Transformation Table*

|  |
| --- |
|  |

### **Transformation on Y**

****

*Figure 20. Transformation on Y*

|  |
| --- |
| *Figure 21. Shapiro-Wilk test* |

Since constant variance and normality cannot be achieved in the model we have established, log transformation is applied on y. When we examine the summary, the number of reviews is still significant for the model. The p-value of the model is still less than 5% and the model is significant. According to the shapiro-wilk test, the model is normally distributed.



*Figure 22. Linear Regression Plot*

However, as can be seen from the plot, the model is not linearly distributed. There are lots of outliers. Therefore, transformation should be applied to the x variable.

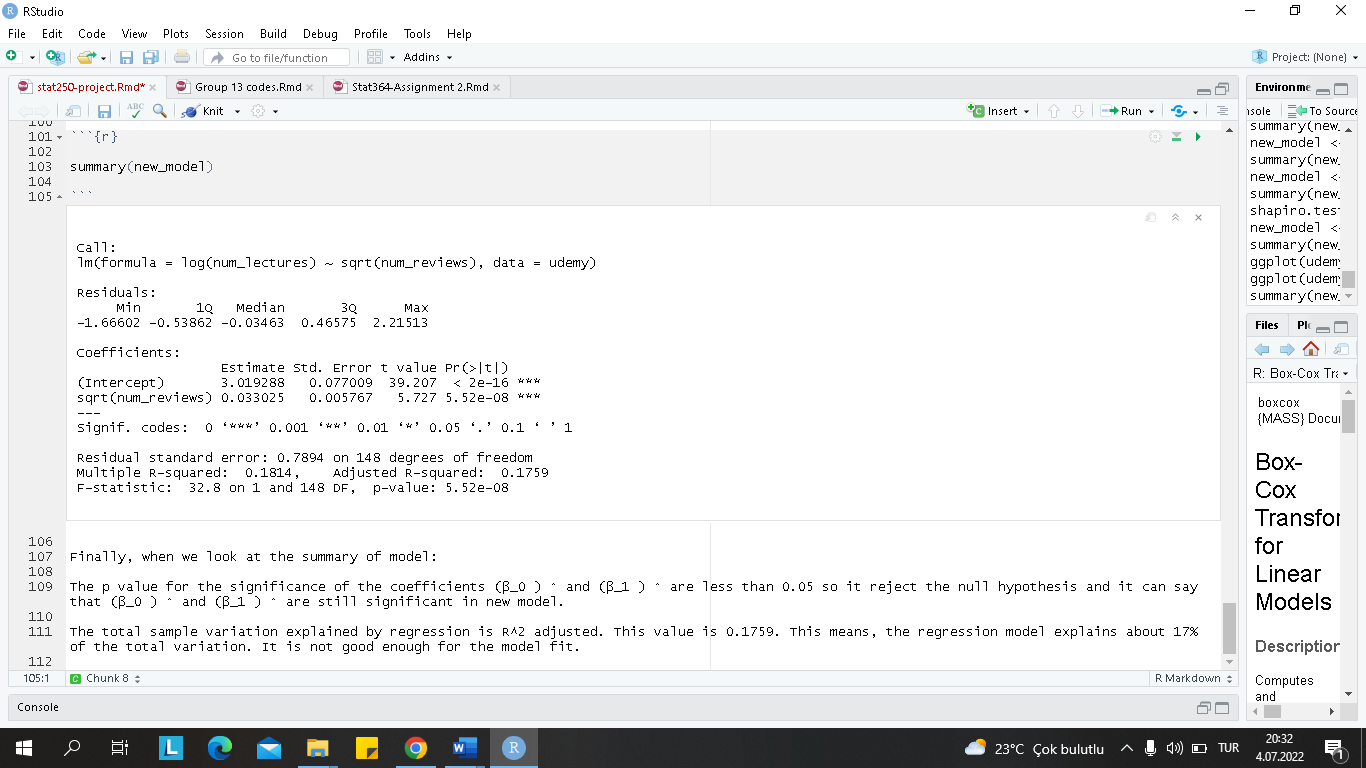
**Transformation on X**



The transformation was also applied to the X (num\_reviews) variable. As can be seen from the scatter plot, the heteroscedasticity problem still persists, but it is now better than the previous model.

|  |
| --- |
| *Figure 23. Scatter Plot for Transformation on X* |

**Transformation Results:**

****

|  |
| --- |
| *Figure 24. Summary for linear model* |

The p value for the significance of the coefficients 0 and 1 are less than 0.05 so it rejects the null hypothesis and it can say that 0̂and 1are still significant in the new model. The total sample variation explained by regression is R-squared adjusted. This value is 0.1759. This means, the regression model explains about 17% of the total variation. It is not good enough for the model fit.

## 

## **3.6 Multiple Linear Regression For Number of lectures, Number of Subscribers, Course Price, Number of Reviews, Content Duration and Money Paid for Course or not**

**Q.** Does number of subscribers, course price, number of reviews, content duration and

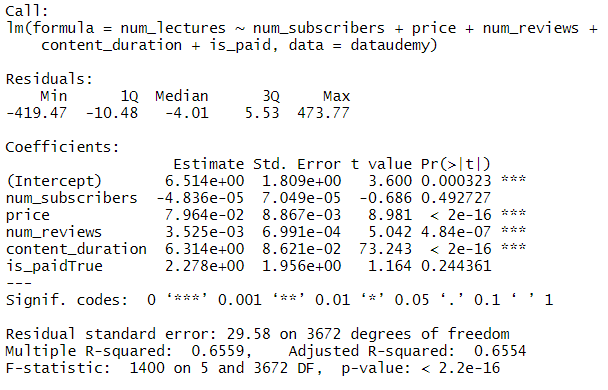
does money paid for courses or not affect the number of lectures?

**Response:** The Number of Lectures

**Proposed method:** Multiple Linear Regression

3.3.1 First Model

Here is the result of the analysis with the proposed method ;



*Figure 25. Multiple Linear Regression Summary*

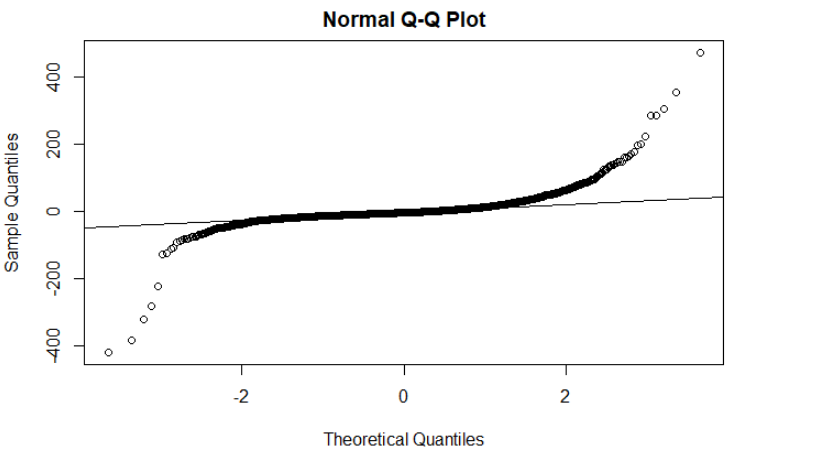
Associated p value of the F-statistics is 2.2e-16, which is highly significant. So, we concluded that at least one of the predictor variables is important and related to the outcome variable. Price, content duration, number of reviews and money paid for course seem to be significant for the model. Let's check the normality of the response with QQ plot and shapiro test.

3.3.2 QQ-Plot and Normality Test

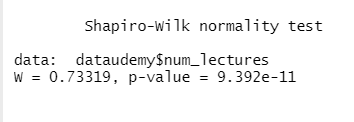
Shapiro-Wilk test checks the normality of the variable. In this approach;

H\_0 :The response is normally distributed

H\_1:The response is not normally distributed



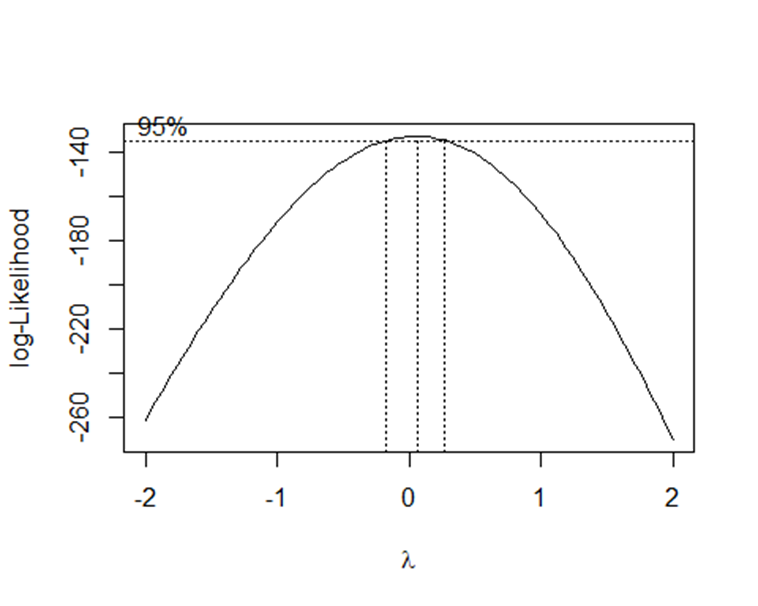
*Figure 26. Normal Q-Q Plot*



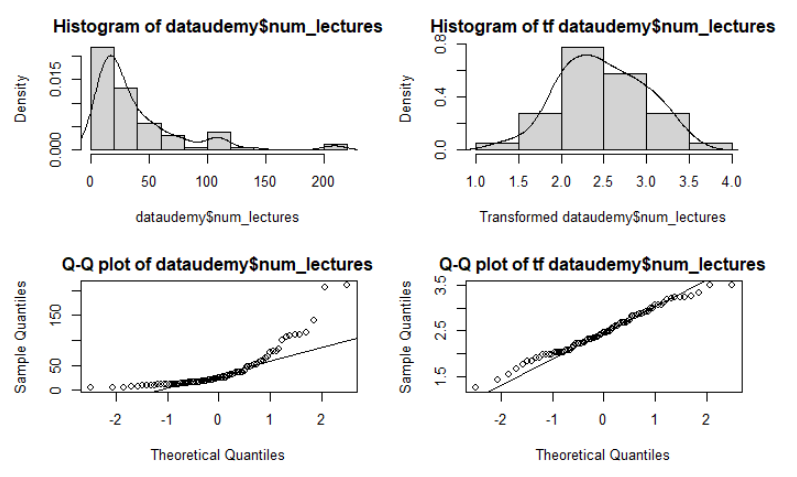
*Figure 27. Shapiro-Wilk Normality Test*

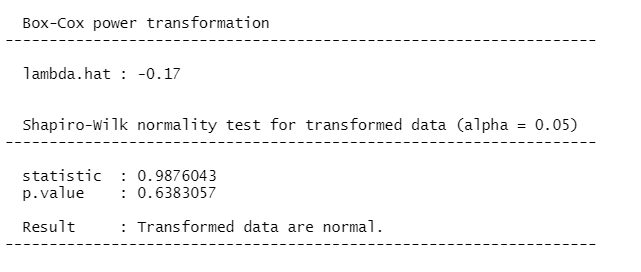
Since p-value is smaller than the alpha value, 0.05, we reject the null hypothesis. Thus, data is not distributed normally. In order to solve non-normality, we should use transformation methods on variables.

3.3.3 Box-cox Transformation

Box-cox transformation is used following;

*Figure 28. Box-cox Transformation Plot*



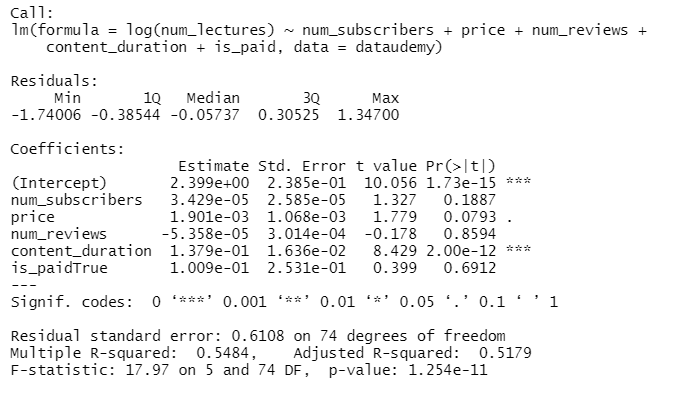


*Figure 29. Box-cox Transformation on Response*

Graph shows at the beginning data is not distributed normally but after transformation, it gets better. It seems to be distributed normally. Also, when we compare to QQ plots, observations are getting close to the line. When we check the box cox plot and see the confidence interval for $\lambda$ includes 0, we may apply $log(y)$ transformation. Since normality and constant variance cannot be provided in the linear model we have established, log transformation is applied on the response variable.

3.3.4 Transformed Model

Here is the transformed model;



*Figure 30. Regression Summary*

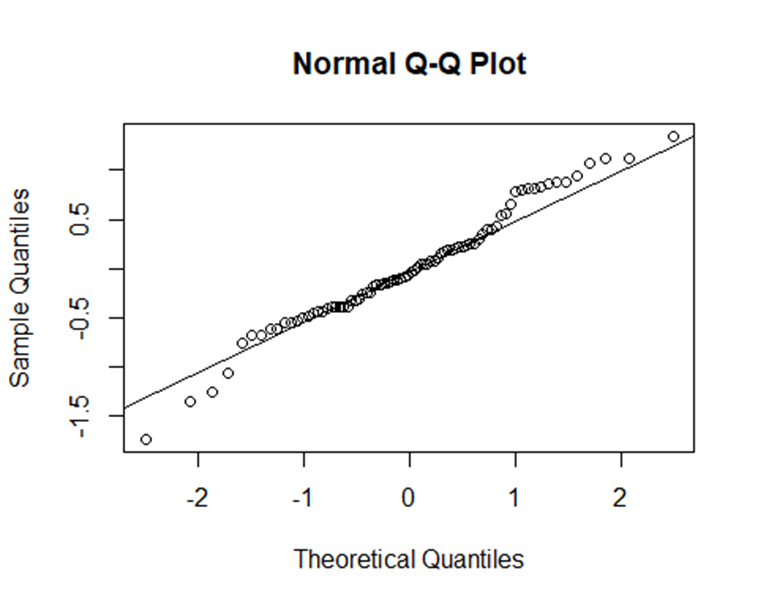
Output shows that at 0.1 significant level, intercept, price and content duration are significant for the model. We remove the variables,number of subscribers (number of reviews and whether money paid for course or not) which are not significant for the model.

3.3.5 QQ-Plot and Normality Test

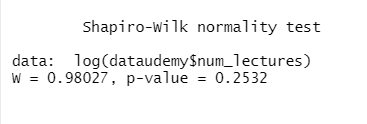
Shapiro-Wilk test checks the normality of the variable. In this approach;

H\_0 :The response is normally distributed

H\_1:The response is not normally distributed



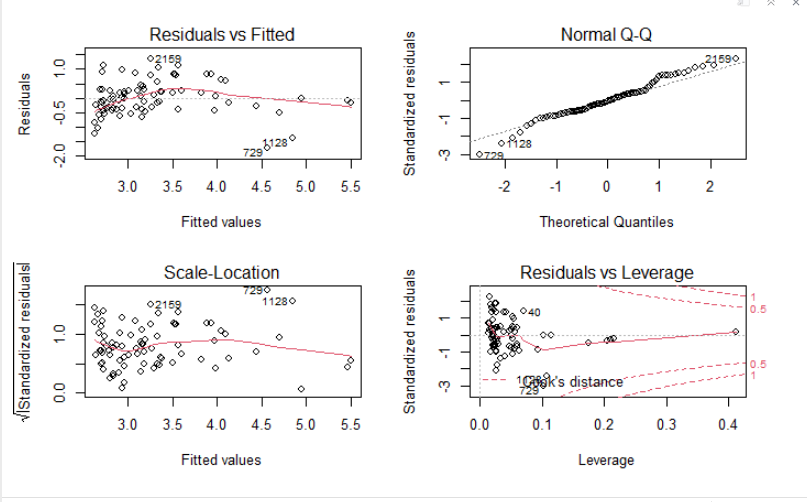
*Figure 31. Normal Q-Q Plot*



*Figure 32. Shapiro-Wilk Normality Test*

QQ plot and normality test which is shapiro test shows that now data is distributed normal after log transformation.

3.3.6 Residuals vs Fitted Values



*Figure 33. Residuals vs Fitted Values*

Residuals are not distributed around zero. There is a pattern. data is not a distributed homogen. According to the chart, there may be residual outliers.

3.3.7 Multicollinearity Check

Variance Inflation Factor

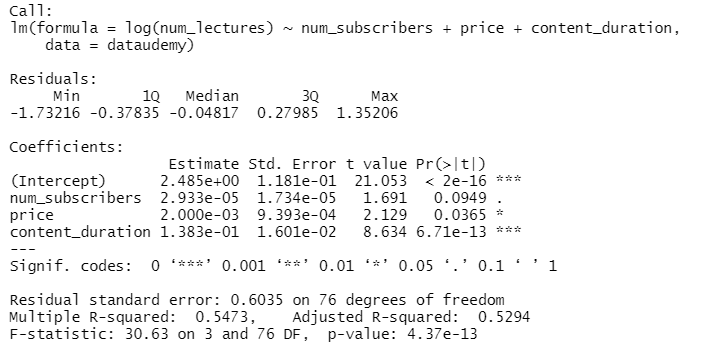


*Figure 34. Variance Inflation Factor Control*

Variance Inflation Factor is smaller than 5 or 10 for number of subscribers, price and content duration. That means there is no multicollinearity problem.

3.3.8 The Best Model

The best version of the model is as below;



*Figure 35. Multiple Regression Summary*

The total sample variation explained by regression is $R^2 adjusted$. It is equal to 0.5473 This means, the regression model explains about 55% of the total variation. It is good enough for the model fit. Model equation is $y= 2.49+ numsubscribers\*0.0000293+ price\*0.002+ contentduration\*0.138$ As we can see from the model equation, one unit change on explanatory variables makes change in response positively.

.

# **4. DISCUSSION/CONCLUSION**

In hypothesis test, tested whether the price of courses and the number of lectures do not have any relations between each other. we concluded that using wilcoxon test.In one sample, tested whether the mean of course price is equal to 40 or not. As a result, we conclude that the mean of course price is different than 40. Also, a hypothesis test is applied for two samples to see if the mean between the number of subscribers and the number of reviews are equal or not. Concluded they are not included. In another test which is one way anova, we try to show whether levels have an effect on prices paid for courses or not. However, since residuals are not normally distributed and each group doesn't have equal variance, we couldn't perform anova test. we couldn't provide assumptions to apply for the anova test. At the beginning we expect that the number of subscribers, and reviews, price, content duration and whether money paid or not affect the number of lectures. First model shows us only content duration is significant for the model. However, there was a normality problem with the data. Normality problem is solved with Box-cox transformation. Also, normality is checked with Shapiro Wil normality test. As a result there is no normality problems on residuals. Residuals vs fitted values plot demonstrates that there is curve on the plot. That means constant error variance problem occurs. There is no homogene distribution on the graph. However we can say that outliers causes this problem. We don't want to remove variables from the data. Without variables residuals line around zero. After transformation and checking variance inflation factor, the best model is created. the latest model says that the number of subscribers, price and content duration have an effect on the number of lectures. Also,The total sample variation explained by regression is equal to 0.5473 This means, the regression model explains about 55% of the total variation. It is good enough for the model fit.We expected the relationship between the number of lectures and the number of views to be positive but weak. It turned out as we expected. We could not obtain a good model because there were too many outliers, it had a great effect on the analysis. Our R-square adjusted value turned out to be quite low. As a result, the model cannot be explained well with the available data.

**5. REFERENCES**

*ANOVA Test in R Programming* (2021, Dec 12). GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/anova-test-in-r-programming/

# *How to Transform Data in R* (Log, Square Root, Cube Root)

(2020, Oct 13). Statology. Retrieved from https://www.statology.org/transform-data-in-r/

*Multiple Linear Regression* (n.d.). Applied Statistics with R. Retrieved from https://book.stat420.org/multiple-linear-regression.html

*Udemy Courses*. Retrieved from https://www.kaggle.com/datasets/andrewmvd/udemy-courses/code?datasetId=646598&language=R