Intermediate Statistics— MOOCs study

Aya Ben Hriz

October 2021

1 Introduction

Few studies have focused on the long-term evolution of the learning engagement patterns in a given course of Massive Open Online Courses (MOOC). To understand ongoing dynamics and to adapt course design accordingly, and to capture ongoing trends at a more global scale, we are going to analyze a MOOC that consist of a entrepreneurship course called ”Effectuation” that last 5 weeks. Our mission is to check whether it's true that most of the time, a large proportion of registrants could still represent a significant part of the course activity despite the fact that they do not complete the course or not.

2 Materials and Methods

2.1 Merging our data sets

We have 7 data sets in total countries, usages1,2,3 and effec1,2,3. We merge effec,1,2,3 and usages1,2,3 to obtain a data frame of 28470 observations and 122 variables and we rename the column HDI with the values ”H”,”M” by ”I” as intermediate to obtain our final data set.

2.2 Describing behavior in the courses

We then categorize participants based on their level of engagement. Those who obtained a certificate were called “completers”, those who submitted at least tone quiz or assignment but did not complete the course were referred to as “disengaging learners”, those who did not submit any quiz or assignment were referred to as “auditing learners” if they had viewed more than 10 percent of available videos, and bystanders if they fell below this threshold.

2.3 Linear model

2.3.1 ANOVA test.

Our goal now is to check using ANOVA which is a statistical test for estimating how a quantitative dependent variable changes according to the levels of one or more categorical independent variables. We want to compare the number of views of videos between genders that means we want to know whether the mean number of videos watched differs according to the gender. We can test the difference between these two groups using a t-test.

2.3.2 One-way ANOVA test.

The test that we have to use is a 2 sample t test since our gender variable has 2 factors. We have 2 options to do that. First one is Welch t-test. We recall that, by default, R computes the Welch t-test, which is the safer one. This is the test where you do not assume that the variance is the same in the two groups, which results in the fractional degrees of freedom. Our second option is checking wether there's a difference between the variances of the 2 sets of data and set the var.equal parameter either to True or to False according to the results. We then compare the number of views of videos depending on the HDI of the country of origin. We run the one-way ANOVA test with the command aov. We use Anova, it's a statistical method which is an extension of the t-test. It is used in a situation where the factor variable has more than one group which is the case of HDI.

2.3.3 Two way ANOVA test.

To compare the number of views depending on Gender and HDI, two-way ANOVA is used since there are more than 1 explanatory variable.

2.3.4 Pairwise comparisons

For this part we update the model, and add an interaction parameter which is Gender*HDI.

We then, perform a stepwise regression (both forward and backward selection) algorithm to assess the performance of various versions of the model.

Our next step is assesing colinearity using a chi-test between HDI and Gender and producing a mosaic plot so that we could interpret the results.

But ANOVA does not tell which groups are different from the others. In order to check the pairwise differences between learners of different socioeconomic status , a post-hoc test must be performed such as Tukey HSD(Tukey Honest Significant Differences).

2.4 Logistic Regression

2.4.1 Producing an odd-Ratios table

We then make use of a logistic regression model, to test whether completion in the course is linked to the user characteristics that we studied earlier and create an Odd-Ratio table. An Odd-Ratio is calculated for a given variable in relation to one of the characteristics of this variable. We will be studying the effect of gender and HDI on the probability of completing the course (a boolean variable taking 1 for completion and 0 otherwise).

2.5 Poisson models for count data

3 Results

3.1 Describing behavior in the courses

Let's recall that the aim of this prospective randomised trial was to assess the impact of the online learning approach on students from different fields and study levels. And that we are expecting that a large proportion of registrants could still represent a significant part of the course activity despite the fact that they do not complete the course.

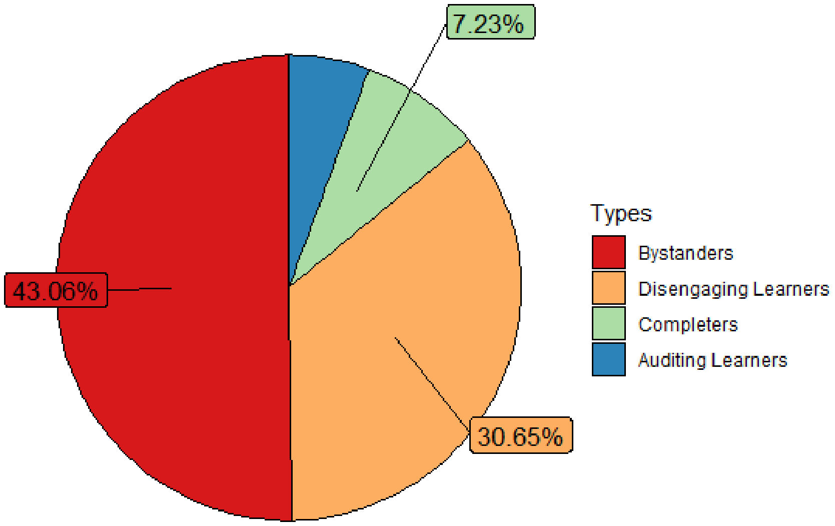
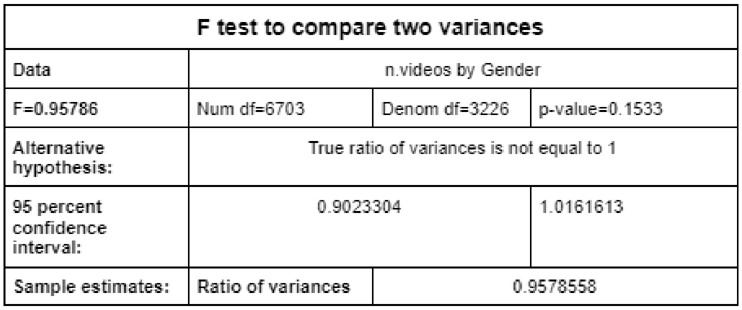


Figure 1: Proportion of learners

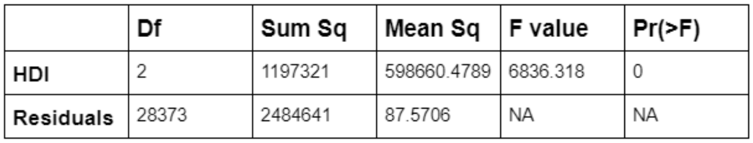
3.2 Linear model

3.2.1 One way ANOVA between number of videos and HDI



F test to compare two variances			
Data	n.videos by Gender		
F=0.95786	Num df=6703	Denom df=3226	p-value=0.1533
Alternative hypothesis:	True ratio of variances is not equal to 1		
95 percent confidence interval:	0.9023304		1.0161613
Sample estimates:	Ratio of variances	0.9578558	

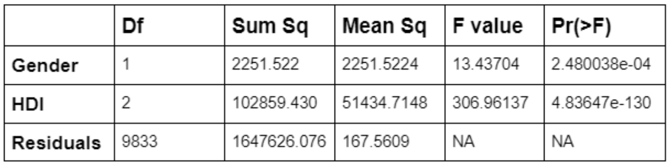
Figure 2: Comparing variances using F test



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HDI	2	1197321	598660.4789	6836.318	0
Residuals	28373	2484641	87.5706	NA	NA

Figure 3: One way ANOVA between number of videos and HDI

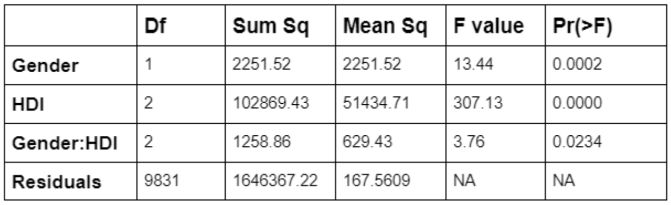
3.2.2 Two way ANOVA with Gender and HDI



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	2251.522	2251.5224	13.43704	2.480038e-04
HDI	2	102859.430	51434.7148	306.96137	4.83647e-130
Residuals	9833	1647626.076	167.5609	NA	NA

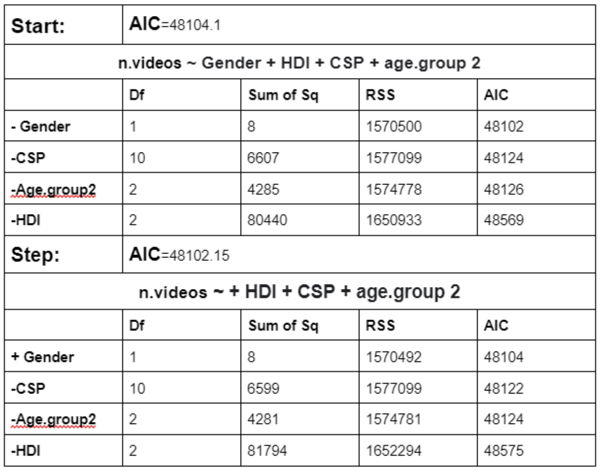
Figure 4: Two way ANOVA with Gender and HDI

3.2.3 Pairwise comparisons



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	2251.52	2251.52	13.44	0.0002
HDI	2	102869.43	51434.71	307.13	0.0000
Gender:HDI	2	1258.86	629.43	3.76	0.0234
Residuals	9831	1646367.22	167.5609	NA	NA

Figure 5: Two way ANOVA (Gender-HDI) with interaction parameter (Gender*HDI)



Start:	AIC=48104.1			
n.videos ~ Gender + HDI + CSP + age.group 2				
	Df	Sum of Sq	RSS	AIC
- Gender	1	8	1570500	48102
-CSP	10	6607	1577099	48124
-Age.group2	2	4285	1574778	48126
-HDI	2	80440	1650933	48569
Step:	AIC=48102.15			
n.videos ~ + HDI + CSP + age.group 2				
	Df	Sum of Sq	RSS	AIC
+ Gender	1	8	1570492	48104
-CSP	10	6599	1577099	48122
-Age.group2	2	4281	1574781	48124
-HDI	2	81794	1652294	48575

Figure 6: StepAIC search method for feature selection.

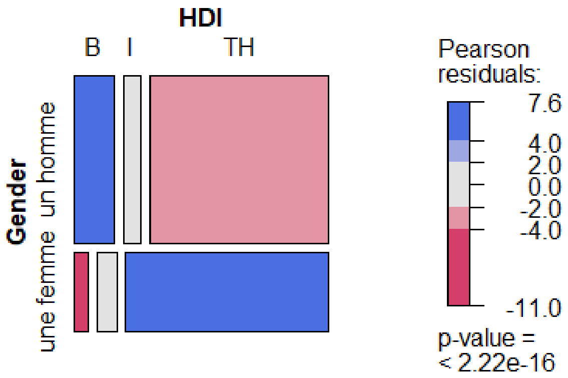
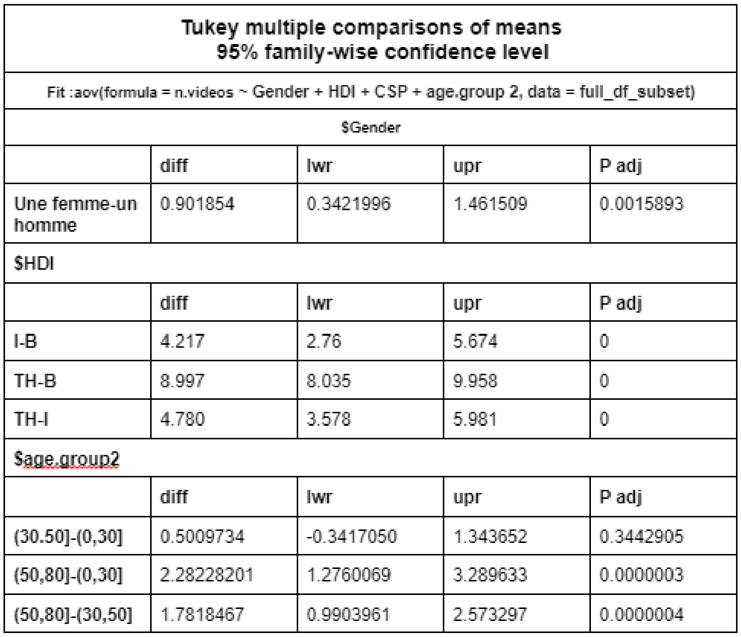


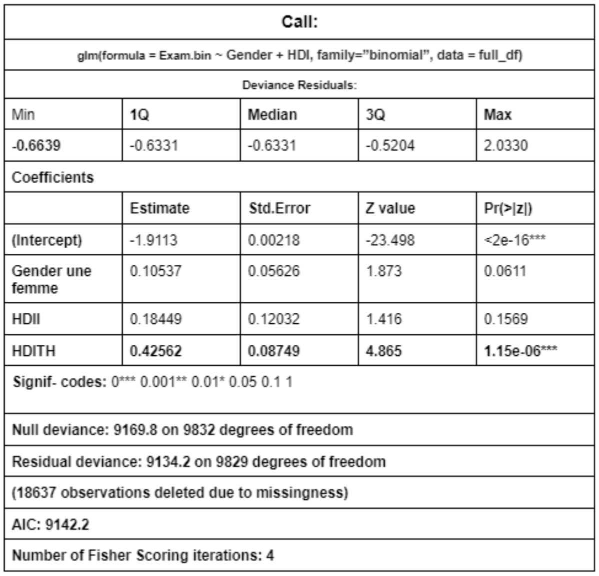
Figure 7: Assesing colinearity



Tukey multiple comparisons of means 95% family-wise confidence level				
Fit :aov(formula = n.videos ~ Gender + HDI + CSP + age.group 2, data = full_df_subset)				
$Gender				
	diff	lwr	upr	P adj
Une femme-un homme	0.901854	0.3421996	1.461509	0.0015893
$HDI				
	diff	lwr	upr	P adj
I-B	4.217	2.76	5.674	0
TH-B	8.997	8.035	9.958	0
TH-I	4.780	3.578	5.981	0
$age.group2				
	diff	lwr	upr	P adj
(30.50]-(0,30]	0.5009734	-0.3417050	1.343652	0.3442905
(50,80]-(0,30]	2.28228201	1.2760069	3.289633	0.0000003
(50,80]-(30,50]	1.7818467	0.9903961	2.573297	0.0000004

Figure 8: Tukey HSD

3.3 Logistic Regression



Call:				
glm(formula = Exam.bin ~ Gender + HDI, family="binomial", data = full_df)				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-0.6639	-0.6331	-0.6331	-0.5204	2.0330
Coefficients				
	Estimate	Std.Error	Z value	Pr(>|z|)
(Intercept)	-1.9113	0.00218	-23.498	<2e-16***
Gender une femme	0.10537	0.05626	1.873	0.0611
HDII	0.18449	0.12032	1.416	0.1569
HDITH	**0.42562**	**0.08749**	**4.865**	**1.15e-06*** **
Signif- codes: 0*** 0.001** 0.01* 0.05 0.1 1				
Null deviance: 9169.8 on 9832 degrees of freedom				
Residual deviance: 9134.2 on 9829 degrees of freedom				
(18637 observations deleted due to missingness)				
AIC: 9142.2				
Number of Fisher Scoring iterations: 4				

Figure 9: Summary of Logistic Regression model

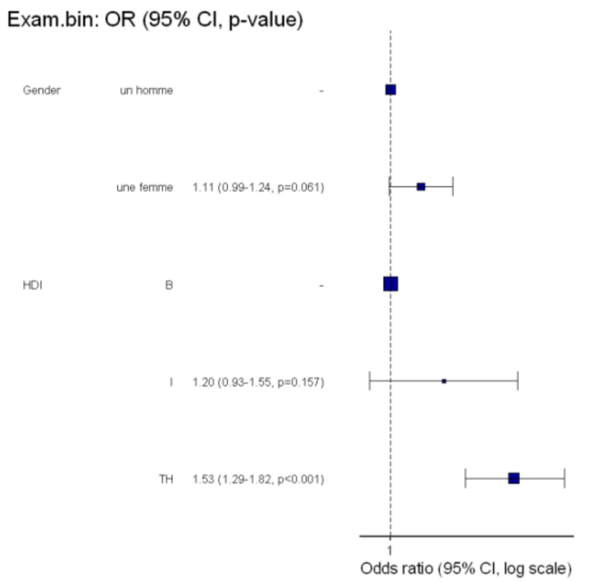


Figure 10: Odd-Ratios plot with 95% confidence intervals and p-values for the influence of gender and HDI on the probability of obtaining a certificate.

3.4 Poisson models for count data

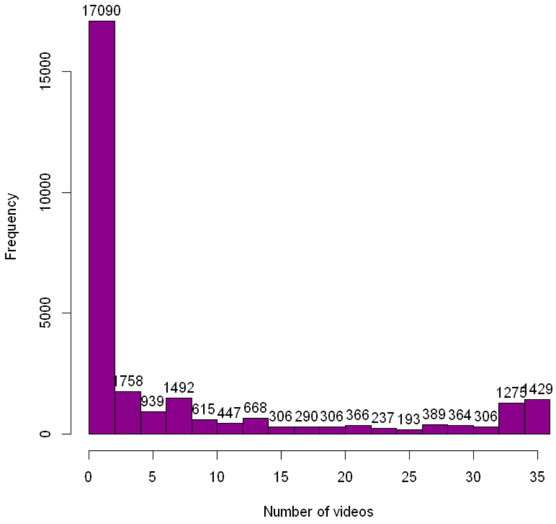


Figure 11: Distribution of the number of videos viewed.

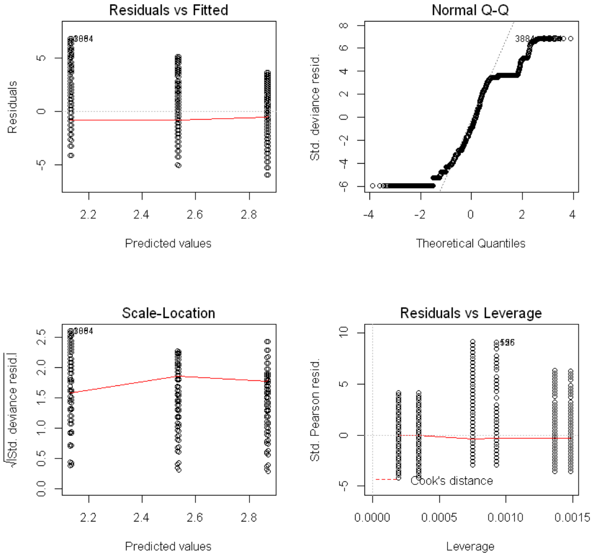


Figure 12: Residual analysis on poisson regression model.

4 Discussion

4.1 Describing behavior in the courses

After categorizing participants on their level of engagement, it turned out that from figure 1, only 7.23% are completers which means they obtained a certificate, 30.65% are disengaging learners and 19.06% are auditing learners. Meanwhile 43.06% are bystanders that did not obtain any certificate neither submitted an assignment nor a quizz and didn't watch at least 10% of the videos. The current research is only limited to theory classes, therefore it can be implemented to check students performance in practical classes. The study is done on students from all over the world; thus,the data is giving comparative results to understand the student's perspective. We may conclude that this MOOC method works only with some specific countries. There may be some issues and problems faced by the students, like the limited access to the internet or disturbance due to low signals. Some of the students may face the home environment issues such as disturbance due to family members, which may lead to negative performance. We also observe a considerable percentage of non existing values which may alter our study and change our conclusions.

4.2 Linear model

4.2.1 One way ANOVA between number of videos and HDI

From Figure 2, the p-value of F-test is p = 0.153. It's greater than the significance level alpha = 0.05. In conclusion, there is no significant difference between the variances of the two sets of data. Therefore, we can use the classic t-test witch assume equality of the two variances.

From Figure 3, F(2,28373) =6836, p = 0.0000, as the p-value is less than the significance level of 0.05, it can be concluded that there is a significant difference between the different groups of HDI.

4.2.2 Two way ANOVA with Gender and HDI.

In figure 4, adding gender has decreased the residual variance from 2484641 (Figure 3) to 1647626 (Figure 4).
From figure 4 for gender, [F(1,9831) = 13.44 , p = 0.0002], there was a statistically main effect on number of videos watched by each gender and for HDI, there was also a statistically main effect on number of videos watched depending on country of origin, [F(2,9831)=307.13 , p =0.0000.]

4.2.3 Pairwise comparisons.

4.2.4 Two way ANOVA (Gender-HDI) with interaction parameter (Gender*HDI).

Furthermore, rather than having gender and HDI to have an additive effect on each other, they may have an interactive effect on each other. In Figure 5, F(2,9831) = 3.76, p = 0.0234, the interaction between gender*HDI is statistically significant, which indicates the relationship between number of videos and gender depend on HDI.

4.2.5 StepAIC search method for feature selection.

Let's recall that our goal is to keep on minimizing the stepAIC value to come up with the final set of features. StepAIC does not necessarily means to improve the model performance, however it is used to simplify the model without impacting much on the performance. In Figure 6, Gender has lowest AIC value 48102 which means the amount of information loss by removing Gender is minimum. Gender will be then removed and stepAIC will run with the remaining set of variables. Plus sign in front of Gender tells that in subsequent iteration, it has also checked by adding the removed variable again if it increases the AIC. At the very last step stepAIC has produced the optimal set of features CSP, age.groupe2, HDI. StepAIC also removes the Multicollinearity if it exists.

4.2.6 Mosaic plot

From figure 7, we notice that the first split is with gender with about 2/3 male and about 1/3 female. The second split is according to HDI (conditional on gender) showing a clear association that females and males tend to get more TH results than B and I. Also men having B and woman TH are the most staticscally significant since the residuals fall between 4.0 and 7.6.

4.2.7 Tukey HSD.

Tukey's multiple comparison test is one of several tests that can be used to determine which means amongst a set of means differ from the rest. When adjusted p-values are less than the significance level, the difference between those group means is statistically significant.
From figure 8, for gender, there is a statistical evidence that females watch on average 0.9 more videos than males (p = 0.0156).
For HDI, there are statistical evidences that countries of different HDIs watch different amount of videos. Countries with medium HDI watch on average 4.21 more videos than low HDI countries(p=0). Countries with very high HDI watch on average 8.99 more videos than low HDI countries(p=0), and very high HDI countries watch on average 4.78 more videos than medium HDI countries(p=0).
For age groups, there is a statistical evidence that age group of 50-80 watch on average 2.28 more videos than the age group of 0-30(p=0.0000003). And for the age group of 50-80, they watch on average 1.78 more than videos than the age group of 30-50(p=0.0000004).
However, there is no statistical evidence to suggest that the age group 30-50 and 0-30 watch different number of videos(p=0.344).

4.3 Logistic Regression.

4.3.1 Summary of Logistic Regression model.

Generalized Linear Models enable the use of linear models in cases where the response variable has an error distribution that is non-normal. Each distribution is associated with a specific canonical link function.
In figure 9, by specifying family = "binomial", glm automatically selects the appropriate canonical link function, which is the logarithm. We see that 'HDII', 'HDITH' and 'Gender, une femme' influences 'Exam.bin' positively. We also see that the coefficient of 'HDII' and 'Gender, une femme' is non-significant (p > 0.05), while the coefficient of 'HDITH' is significant.
We see the word deviance twice over in the model output. Deviance is a measure of goodness of fit of a generalized linear model. Or rather, it's a measure of badness of fit–higher numbers indicate worse fit. R reports two forms of deviance – the null deviance and the residual deviance. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean). For our example, we have a value of 9169.8 on 9832 degrees of freedom. Including the independent variables (weight and displacement) decreased the deviance to 9134.2 points on 9829 degrees of freedom, a significant reduction in deviance.

The Residual Deviance has reduced by 35.6 with a loss of three degrees of freedom. What about the Fisher scoring algorithm? Fisher's scoring algorithm is a derivative of Newton's method for solving maximum likelihood problems numerically.

For model1 we see that Fisher's Scoring Algorithm needed 4 iterations to perform the fit.

4.3.2 Odd-Ratio plot

From Figure 10, for gender an odd-ratio value of 1.11, signifies that females are 1.11 more likely than men to complete the course. It does not mean that there is a 1.11 ratio between the completion rate of women to that of men. However, with each odd-ratio value, there is a p-value associated to determine the statistical signficance of the odd-ratio value. A star means that the p-value is less than 0.05, which means that the test is statistically significant with a margin of error of 5%. Two stars mean that the p-value is less than 0.01 and that the test is statistically significant with a margin of error of 1%, while three stars means that the p-value is less than 0.001. The lower this value the more statistically significant is the difference between the groups. So this means that there is no statistical evidence that women are 1.11 more likely than men to complete the course ($p = 0.061$). For HDI, only one odd-ratio of out two is statistically significant, countries with very high HDI are 1.53 times more likely than low HDI countries to complete the course ($p < 0.001$).

4.4 Poisson models for count data

4.4.1 distribution of the variable (videos viewed).

We notice from Figure 14, that the mean is greater than the median since the shape of the distribution is positively skewed. This positive skewness means that for a smaller number of videos, there are more views and for the higher number of videos, there are less views. The distribution is indeed not normal and therefore follows a poisson distribution.

4.4.2 Residual analysis on poisson regression model.

In figure 15 for Fitted vs Residuals plot, we see that linearity seems to hold reasonably well, as the red line is close to the dashed line.
As for Residuals vs Leverage plot, we're looking at how the spread of standardized residuals changes as the leverage, or sensitivity of the fitted $\hat{Y}$ to a change in yi, increases. Firstly, this can also be used to detect heteroskedasticity and non-linearity. The spread of standardized residuals shouldn't change as a function of leverage: here it appears to decrease, indicating heteroskedasticity. Second, points with high leverage may be influential: that is, deleting them would change the model a lot. For this we can look at Cook's distance, which measures the effect of deleting a point on the combined parameter vector. Cook's distance is the dotted red line here, and points outside the dotted line have high influence.

For Normal QQplot, the points approximately fall on the line, but what does this mean? The simplest explanation is as follows: say you have some observations and you want to check if they come from a normal distribution. You can standardize them (mean center and scale variance to 1) and then 'percentile match' against a standard normal distribution. Then you can plot your points against a perfectly percentile-matched line. Now how can we characterize the (slight) non-normality? What we see is that on the right hand side of the graph, the points lie slightly above the line. For the very right-most point, this is saying that the value x such that P(X)=0.99 is larger under the empirical CDF for the standardized residuals than it is under a normal distribution. This suggests a 'fat tail' on the right hand side of the distribution.

The scale-location plot is very similar to residuals vs fitted, but simplifies analysis of the homoscedasticity assumption. It takes the square root of the absolute value of standardized residuals instead of plotting the residuals themselves. Recall that homoscedasticity means constant variance in linear regression. The red line is approximately horizontal. Then the average magnitude of the standardized residuals isn't changing much as a function of the fitted values.

5 Survival Analysis

6 Conclusion

The future research can use a longitudinal study to handle this limitation. Further, the data was collected from one type of respondents only, that is, the students. Therefore, the results of the study cannot be generalized to other samples. The future research can also include the perspectives of teachers and policy makers to have more generalization of the results.