Statistical Case-Study on Learners Engagements in MOOCs

Amr MOHAMED
CY Tech/ CY Cergy Paris University
mohamedamr@cy-tech.fr

Abstract

In this paper, we are going to go through a statistical case study that was done over a real-life dataset of learners' engagement behavior in MOOCs. In the study, we will see how the learners' genders, countries' Human Development Index (HDI), and socioeconomic status can affect the engagement behavior of the learners. We found that users who engaged the most in the MOOC were learners from countries with very high HDI.

1 Introduction

Over the past decade, online learning and MOOCs became a hot topic on which there is a debate whether the MOOCs have good quality education and if in the future, MOOCs can replace universities education. in this paper we are going to have a close look at the learners' engagement behavior in the "Effectuation: l'entrepreneuriat pour tous" MOOC [1] by Professor Philippe Silberzahn, EMLYON Business School. In addition, we are going to study the completion rates of the course, and discover the effect of the learners' gender, countries, and socioeconomic status on their likely hood of engaging in the MOOC and completing it.

2 Methods

2.1 Data wrangling, feature engineering

The data that was used in this study was gathered over 3 enrollment cycles and for each cycle a data 2 files. In addition, a Human Development Index by countries data file was provided. The overall data structure after merging all the files is 120 columns and 9760 rows.

Firstly, all the files in section 2.1 were joined by rbind.fill() and cbind.fill(). Afterward, 62 columns were selected after prioritizing the importance of the columns and viewing the missingness in the data, where the columns which had more than approximately 60% of their data missing and can't be retrieved from other columns were dropped out.

In Addition, some data required some modifications, for example, the users from Canada were divided into two different categories by location: English-speaking cities and french speaking cities, those two categories were grouped in one category called Canada. Moreover, the birth. year column was a mix of birth years and age of the learners, all these entries were set to the age of the learners.

Afterward, 3 new columns were added indicating the engagement score (the sum of all binary scores of the engagement actions like watching a video or taking a quiz), the engagement percentage, and the engagement category whether the learner is a completer who obtained a certificate), a disengager who submitted at least one quiz or assignment but did not complete the course, an auditor who did not submit any quiz or assignment, but they had viewed more than 10 percent of available videos, or a bystander who did not meet any of the criterion mentioned above.

3 Results

3.1 Describing behavior in the courses

Firstly, we wanted to know the countries where the users enrolled in the MOOC were from, so we started by plotting a bar graph for the top 10 countries that have the highest number of users enrolled in the MOOC.

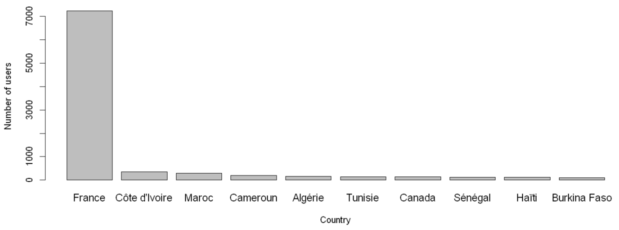


Figure 1: Countries with the greatest number of users enrolled in the MOOC.

As we can see in Figure 1, the countries from which the learners were enrolled in the MOOC were mostly the French speaking countries, with France as the first

[1] https://www.coursera.org/learn/effectuation

country with a big difference from the second country which is Côte d'Ivoire.

Afterward, It was interesting to find the number of learners of each learner engagement behavior, so we retrieved the number of learners of each behavior as in Table 1 below.

Learner engagement behaviour	**Number of learners**
Completer	1708
Disengager	8425
Auditor	774
Bystander	6149

Table 1: Number of learners by engagement behaviour

Moreover, we were interested to see the percentage of the different engagement categories of the learners, so we plotted a pie chart to see the ratio between the different categories.

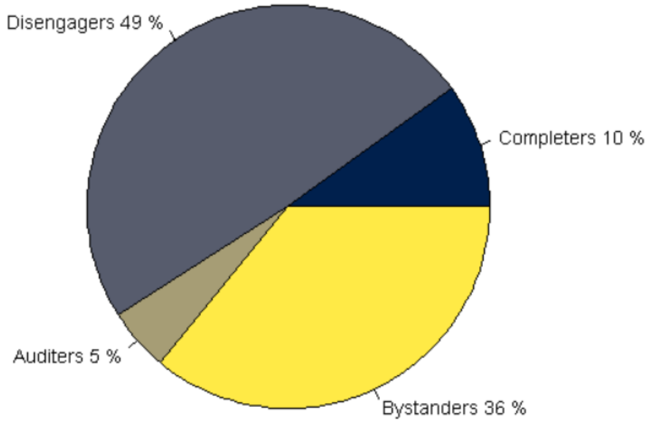


Figure 2: Learners engagement behavior ratios by engagement categories.

As we can see from Table 1 and Figure 2, 49% of the enrolled learners disengaged in the MOOC with several learners of 8425 learners, followed by 36% of bystanders with many learners of 6149, 10% of completers with several learners of 1708 learners, and 5% of auditors with many learners of 774 learners.

3.2 Linear model

3.2.1 From Student's t-test to three-ways ANOVAs

Afterward, we were interested in comparing the number of views of videos between genders, therefore we performed a Student's t-test as shown in table 2 below.

	t
Test statistic	-3.37
DF	5575
p value	0.001
Alternative hypothesis	two.sided

Table 2: Welch's Two Sample t-test of Number of videos by Gender

As we can see from Table 2, we reject the null hypothesis of the test which is that the true difference in the mean number of videos watched between male and female learners is equal to 0 with a 95 percent confidence interval, and we have sufficient evidence to say that the mean number of views of videos for the populations of both genders are different.

Moreover, it was interesting to know the effect of the learner's country Human Development Index (HDI) on the learners' engagement behavior, so we performed a one-way ANOVA test on the HDI category (very high, intermediate, or low) since we want to know if there are significant differences between the means of the different groups of the independent variable HDI by the dependent variable number of videos watched. Here we used a one-way ANOVA test as we have here 3 categories of the HDI variable that's why we can't use a t-test here as we did for the t-test from Table 2 where we were comparing the statistical significance between the means of the number of videos watched by each of the 2 genders.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HDI	2	76126	38063	225	0
Residuals	8155	1380643	169		

Table 3: One-way ANOVA test of Number of videos by HDI

From Table 3, we can conclude that there are significant differences between the different groups of HDI in the mean number of watched videos(p-value = 0).

Afterward, it was interesting to gather both the gender and HDI variables along with the socioeconomic status of the learner (CSP) to see the differences in the mean number of videos watched by the learners by each category of each variable.

	Estim.	Std. Err.	t val.	Pr(>|t|)
(Intercept)	5.2	4.9	1.1	0.29
Gender:				
une femme	0.12	0.31	0.4	0.69
HDI:				
TH	9.57	0.49	19.5	0
I	4.58	0.71	6.44	0
CSP:				
Artisans	2.13	4.95	0.43	0.68
Prof. intellect.	3.72	4.9	0.75	0.45
Employés	3.35	4.94	0.67	0.5
En rech. d'emploi	4.86	4.93	1	0.3
Etudiants	2.6	4.93	0.53	0.6
Inactif	5.6	5.1	1.10	0.27
Ouvriers	5.26	5.75	0.91	0.36
Prof. interméd.	1.8	5	0.36	0.72
Retraités	4.39	5.2	0.84	0.39

Table 4: Linear model of the Gender, HDI, and CSP variables

From Table 4, we can see that on average female learners have watched 0.12 more videos than male learners. In addition, we can see that learners from very highly developed countries (HDI TH) watched on average 9.57 more videos than learners from low developed countries (HDI B), while learners from Intermediately developed countries (HDI I) watched on average 4.58 more videos than learners from low developed countries. Moreover, we can see that the learners with the highest number of watched videos had a socioeconomic status of inactive, workers, or looking for a job.

Afterward, we wanted to know how the mean number of videos watched by the learners depends on the categorical variables used in the model from the table, Gender, HDI, and CSP, so we performed a three-way ANOVA test on these variables.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	1651	1651	10	0.002
HDI	2	74514	37257	221	0
CSP	10	6065	606	4	0.0001
Resid.	8144	1374538	169		

Table 5: Three-way ANOVA test of Number of videos by HDI, Gender, and CSP

From Table 5, we can see that:

- Both of the genders of learners have a different mean number of videos watched by learners of each gender (p-value = 0.002).
- Learners from countries with different HDI have a different mean number of videos watched(p-value = 0).
- learners with different CSP categories have a different mean number of videos watched.(p-value = 0.0001).

3.2.2 Model refinement, pairwise comparisons

Afterward, we decided to use the linear model used earlier in Table 4, but this time we added an interaction parameter between the gender of the learner and his country's HDI to see whether there is a significant relationship between those variables, and how it can affect the number of videos watched by the learner.

	Estim	Std. Err	t value	Pr(>|t|)
(Intercept)	5.18	4.92	1.1	0.29
Gender:				
une femme	0.98	1.28	0.77	0.44
HDI:				
TH	9.66	0.53	18.1	0
I	5.3	0.83	6.38	0
CSP:				
Artisans	2.05	4.95	0.41	0.68
Prof. intellectuelles	3.65	4.93	0.74	0.46
Employés	3.23	4.94	0.65	0.51
En rech. d'emploi	4.8	4.94	0.97	0.33
Etudiants	2.54	4.93	0.52	0.60
Inactif	5.53	5.1	1.1	0.28
Ouvriers	5.21	5.75	0.91	0.36
Prof. intermédiaires	1.71	5	0.34	0.73
Retraités	4.29	5.21	0.82	0.41
Gender*HDI:				
une femme: TH	-0.77	1.32	-0.58	0.56
une femme: I	-2.6	1.72	-1.51	0.13

Table 6: Multivariate linear regression of Number of videos by HDI, Gender, and CSP with the interaction parameter Gender*HDI

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	1651	1651	9.7	0.002
HDI	2	74514	37257	220	0
CSP	10	6065	606	3.6	0.0001
Gender: HDI	2	446	223	1.3	0.2682
Residuals	8152	1380157	169		

Table 7: Three-way ANOVA test of Number of videos by HDI, Gender, and CSP with the interaction parameter Gender*HDI

From the linear model in Table 6 and the ANOVA test in Table 7, we don't see a statistical significance when adding the interaction parameter of Gender*HDI to the linear model (p-value = 0.27 and F-value = 1.3). In addition, we can see that there is a high ratio between the mean number of videos watched by each gender.

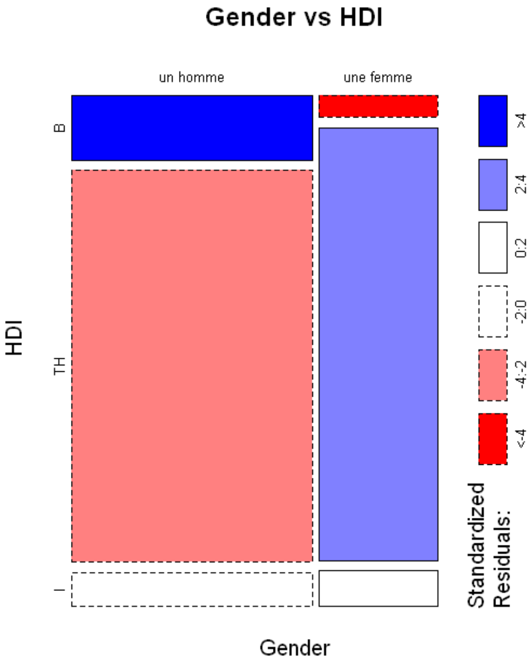
By applying step() function in R for backward, forward, and both directions, we got the results in the following table:

Linear Model	AIC
n.videos ~Gender+HDI+CSP+Gender*HDI	41966
n.videos+ HDI + Gender:HDI	41912
n.videos+ Gender + HDI	41897
n.videos~HDI	41888

Table 8: AIC values for both directions step function on the linear model

As represented in Table 8, all the different combinations of the model parameters perform approximately the same, and the model of only the HDI variable by the number of videos is the best in fitting the data according to the step function.

Afterward, to assess the colinearity of the independent variables of the model used in Table 6, we used the chi-squared test on the HDI and Gender variables and found that there is a significant very high correlation between the two variables (p-value $< 2.2e-16$).



A

Figure 3: Mosaic plot of HDI vs Gender

	B	TH	I
un homme	8.84%	53.53%	4.61%
une femme	1.47%	29.14%	2.41%

Table 9: Contingency table of the percentages of HDI v. Gender

From Figure 3 and Table 9, we can see that that male learner from low developed countries was found in the dataset more than what was expected (standardized residuals $>$ 4) while male learners from very highly developed countries were found less than what has been expected in the dataset(standardized residuals are between -4 and -2), and male learners from intermediately developed countries were found approximately equal to the expected number of learners in the dataset from both categories (standardized residuals are between -2 and 2). Moreover, for female learners, we can see that those of them who are from low developed countries were found strictly fewer than what was expected in the dataset (standardized residuals $>$ 4) while those from very highly developed countries were found slightly more than what was expected (standardized residuals are between 2 and 4), and from intermediately developed countries were found approximately equal to the expected number of learners in the dataset from both categories (standardized residuals are between -2 and 2).

Moreover, we performed Tukey HSD on the CSP variable to see the pairwise differences between learners of different socioeconomic status (CSP), and found the biggest statistically significant pairwise difference in the number of videos watched between learners who are inactively employed Inactive (other than students, retired, or looking for a job) and learners who are employees with a difference of 4.6 videos watched from the MOOC (p-value < 0.0001) while the lowest statistically significant pairwise difference in the number of videos watched between learners who are employees and learners who are managers or in intellectual professions (p-value < 0.05).

3.3 Logistic Regression

3.3.1 Producing an Odd-Ratios table

Afterward, we were interested in studying the different categories of the learners who finished the MOOC, and know what sectors of the learners were more likely to complete the MOOC, so we performed a logistic regression by taking the variables HDI and Gender as the independent variables and the variable (exam.bin), which is a boolean of whether the learner completed the MOOC or not, as the dependent variable.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.73	0.1	-18	0
Gender femme	0.13	0.06	2.2	0.026
HDI TH	0.27	0.10	2.7	0.01
HDI I	0.14	0.15	0.95	0.3

Table 10: Logistic regression of the MOOC's completion by Gender and HDI

From Table 10, we can see that, on average, male learners completed the MOOC with a ratio of 0.13 more than male learners, while learners from countries with low development index, had a lower-rated of course completion by a ratio of 0.27 from learners from very highly developed countries and a ratio of 0.14 than learners from intermediately developed countries.

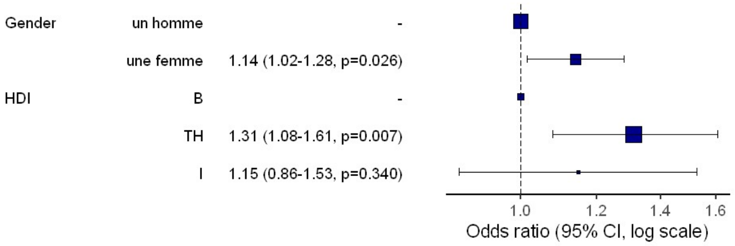


Figure 4: Odd Ratios of MOOC completion by Gender and HDI

From Figure 4, we can see that women had a higher chance of completing the course than men, while the number of women in the study was smaller than the number of men, which means that women were more likely to finish the MOOC than men. For HDI, (we can see a statistically significant p-value for TH), which means that learners from countries with very high HDI were more likely to finish the MOOC.

3.3.2 Poisson models for count data

We were interested in knowing how does the variable number of videos watched by the learners is distributed, so we plotted a histogram as in the Figure below.

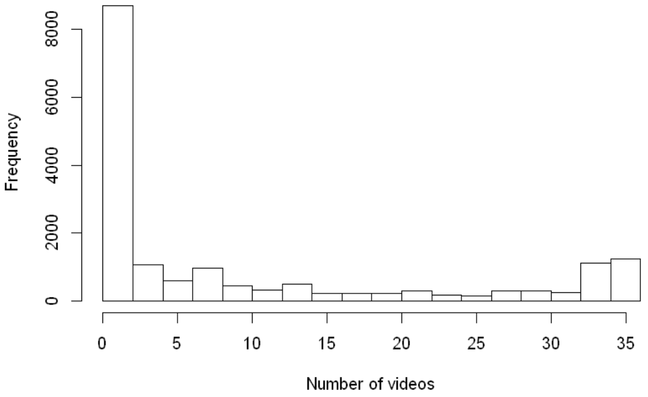


Figure 5: Distribution of the number of videos watched by learners

As we can see in Figure 5, The distribution doesn't follow a normal distribution, and it has 1 very high peak at zero videos, one at zero videos, and another at 35 videos.

To verify our observation about the data distribution, we plotted a QQ-plot.

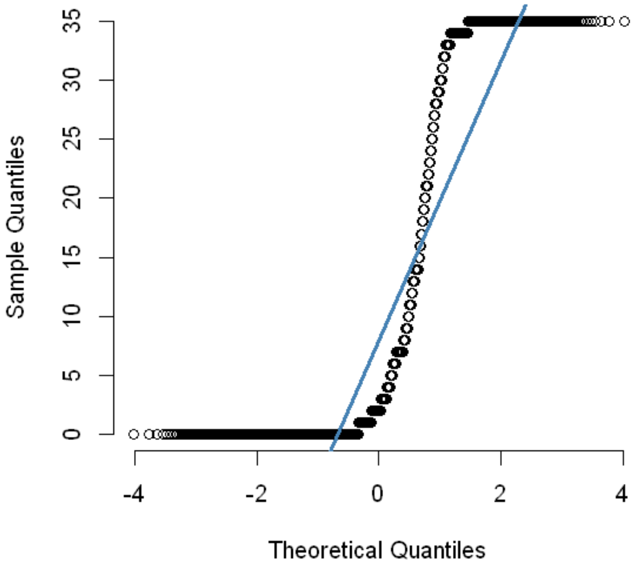


Figure 6: QQ-plot of number of videos watched by learners

As we can see in Figure 6, the data doesn't fit the line and our assumption about the distribution of the data is verified.

Afterward, we decided to create a Poisson Regression model to model the number of videos watched by learners of each gender and each HDI category as the number of videos watched is discrete data (counts) with non-negative values.

	Estim.	Std. Err.	z val.	Pr(>|z|)
(Intercept)	2.17	0.01	185.9	0
Gender une femme	0.01	0.005	1.51	0.13
HDI TH	0.73	0.01	60.8	0
HDI I	0.41	0.02	25	0

Table 11: Poisson Regression model of the number of videos by gender and HDI

From Table 11, we can see that on average female learners have watched 1.01 more videos than male learners (exp(0.01), we take the exponential of the coefficients as the coefficients of the Poisson regression model are log-transformed and we have to exponentiate them to get the baseline coefficients) while for the HDI categories, we can see that learners from very highly developed countries (HDI TH) watched on aver- age 2.1 more videos than learners from low developed countries (HDI B), while learners from Intermediately developed countries (HDI I) watched on average 1.5 more videos than learners from low developed countries (HDI categories coefficients are statistically significant with P-value = 0 for each category).

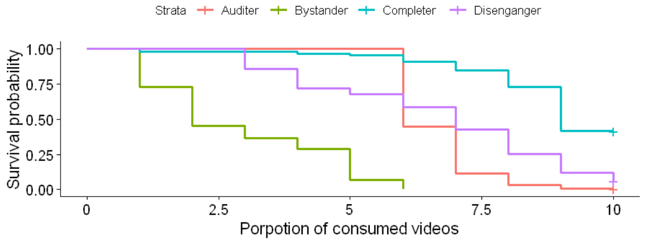
3.4 Survival analysis

Furthermore, we were interested in analyzing the expected proportion of videos that users with different engagement behavior consume before finishing the course or dropping out and failing to complete the course.

Therefore, we performed a survival analysis using The Kaplan–Meier estimator method which involves computing probabilities of occurrence of an event at a certain point in time. These successive probabilities are multiplied by any earlier computed probabilities to get the final estimate. The Kaplan-Meier estimator method uses the following function to calculate the survival probability at any given time:

$$S_t = \frac{\text{Number of subjects living} - \text{Number of subjects died}}{\text{Number of subjects living}}$$

We performed the survival analysis by taking the time variable as the number of videos decile in which the learner has finished or dropped out from the MOOC.



The p

Figure 7: Proportion of videos consumed by learners with different engagement behavior types

As we can see in Figure 7, we can see that completer learners' survival probability was approximately equal to 1 till the 4^{th} decile then it started declining slowly over time till it reached approximately 0.75 survival probability at the 8^{th} decile and had its largest drop at the 9^{th} decile to reach 0.5 survival probability by the end of the MOOC. While for Disengager learners survival probability was approximately 1 from the start of the MOOC till the 3^{rd} decile, then it started declining continuously with relatively small declination steps till it reached approximately 10% by the end of the MOOC. For Auditor learners, surprisingly, they have sustained a survival probability of 1 between the start of the MOOC and the 5^{th} decile, but they suffered from 2 severe drops in the survival probability of 0.5 and 0.35 in the 6^{th} and 7^{th} deciles respectively and then continued in the taking small declination steps till their survival probability reached 0 by the end of the MOOC. And for bystander learners, they had the first declination step after the first decile to reach 0.75 survival probability and continued in taking irregular declination steps till their survival probability reached 0 at the 6^{th} decile.

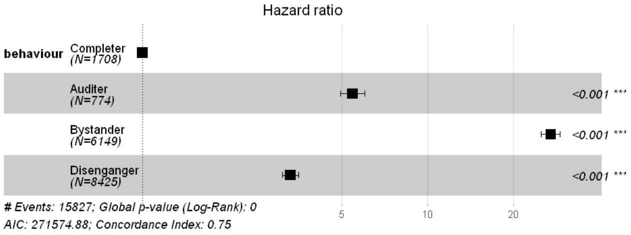


Figure 8: Hazard ratio graph of learners by the different engagement behaviors

From Figure 8, we can see that on average, completers tended to complete the MOOC approximately 3 times faster than disengagers, 6 times faster than auditors, and unsurprisingly, 25 times faster than bystanders.

4 Discussion

We found that learners from very highly developed countries had the highest number of watched videos and after this finding further more verified when we found that those learners had the highest completion rates of the MOOC and we can say now that this is as a result of that those learners have easier access to the internet and online education while on the other side, learners from low developed countries had the lowest completion rates because usually, they have difficulty accessing to internet.

Moreover, we found significant evidence that the socioeconomic status of the learners affected their engagement behavior by affecting the number of videos watched by each category of them, and this may be as a result of that the learners from different socioeconomic statuses have different amounts of spare time that they can allocate to watch the videos of the MOOC, and it also may be because of the difference in priorities between the learner from the different socioeconomic status.

In addition, when adding the interaction parameter between the gender of the learner and his/her country's HDI, we didn't find a significant change in the number of videos watched per gender as the HDI category changes, and by checking the proportions of learners by their gender and country's HDI, we found that most of the learners who took the MOOC where Males from very highly developed countries followed by females from the same HDI category, and the least category that engaged in the MOOC was females who are from low developed countries. And we suggest that these results are because very highly developed countries are a well-prepared place for entrepreneurs to start their businesses while it's harder to start your own business in less developed countries, and that's why we can see that a lower proportion of learners from the lower developed countries enrolled in the MOOC.

Furthermore, we found that the best variable to use when aiming to model the number of videos watched by the learner with a linear model is the HDI variable.

Finally, we found for the learner engagement types that completers were the fastest to finish the MOOC with a probability of 0.5 by the end of the MOOC, and that's because completers most probably finish all the videos and quizzes, and take the final exam while bystanders were the slowest and the least likely to complete the MOOC as they often fall before the 10% thresh-hold of the MOOC. We observed also that auditors had severe drops in the probability of finishing the MOOC as they were going through the MOOC, and this is because auditors didn't do any quizzes, and they were only going through the videos of the MOOC.

References

Magdalena Szumilas (August 2010). ”Explaining Odds Ratios”. :https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757

Anesth Analg (September 2018). ”Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare”. :https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6110618/

Anesth Analg (September 2018). ”POISSON REGRESSION”. :https://stats.oarc.ucla.edu/r/dae/poisson-regression/

Spotswood L. Spruance, Julia E. Reid, Michael Grace, and Matthew Samore (August 2004). ”Hazard Ratio in Clinical Trials”. :https://www.ncbi.nlm.nih.gov/pmc/articles/PMC478551/

Ashutosh Tripathi (June 2019). ”What is stepAIC in R?”: https://ashutoshtr.medium.com/what-is-stepaic-in-r-a65b71c9eeba