# Analysing Factors Correlated to One's Reading Habit

KD7488

21/12/2020

## Abstract

Reading is one of the hobbies that can influence one's life permanently. However, as the technology develops and with the appearance of many other social medias, people read less. In this paper, the main objective is to find out factors that are potentially influential in determining one's reading hobbbies. To accomplish that, a multiple linear regression model is fitted. As a result, people tends to read paper books more and with higher educational background, people read more books in a year.

## Keywords

*Reading Habit*, *educational level*, *age*, *income group*, *How many books one reads in a year*

## I. Introduction

Gorky once said: "Books are the stepping stones to human progress (L,1)". Reading has always been promoted as a positive behavior to human beings, as it is a tool for people to gain knowledge, and pass this knowledge on to generations. Regardless of the media people read from, people can always gain knowledge from books, young or old. However, as people nowadays have many entertainment methods to spend time on, even books have changed from paper to more convenient ways like digital books, fewer and fewer people spend time on readings. Though many potential factors can be correlated to one's reading habit: age, race, marital status, income, etc. For example, people of different age groups have different life routines, which may affect how many books they read or can read in a day.

In this research paper, we are interested in building a linear regression model, where the reading habit of one person is the response variable, and it is measured as the number of books one reads in a year. This study will be conducted based on the dataset provided by Vipul on the website *Kaggle* and the original dataset is collected through *Pew Research Center*.

The paper contains five main parts. Part II focuses on data cleaning and analysis. Part III explains the model we chose, including benefits, drawbacks, and justification. In part IV, the result of the paper will be thoroughly discussed. Finally, part V will be discussing the study results and give interpretations of the model. Lastly, part VI delivers the weakness of the model, including reflections and the next steps to have a final valid model.

The relevant codes and the report in pdf format can be found in the following link:

## Methodology

## II. Data analysis

As the model focuses on the number of books one reads in a year, the variables that suit to be a regressor for the model are *Age*,*Education*,*Incomes*,*Employment*,*How.many.books.did.you.read.during.last.12months.*,*Read.any.audiobooks.*

The variables have rather long names and can be difficult to read, for simplicity they are renamedd in the cleaning data process as *age*,*education*,*income*,*employment*,*num_books*,*read_audio*,*read_ebook*,*read_print*. The structure of the data is presented in the table below:

| variable_name | meaning | min | max | mean |
|---|---|---|---|---|
| age | the age of the individual | 16 | 93 | 45.4 |
| education | the educational background group one belongs to | 0 | 0 | 0.0 |
| income | the income group one belongs to | 0 | 0 | 0.0 |
| employment | one's employment status | 0 | 0 | 0.0 |
| num_books | the number of books individual reads for the past year | 1 | 97 | 19.4 |
| read_audio | whether one listens to audio books for the past year or not | 0 | 1 | 0.2 |
| read_ebook | whether one reads ebooks for the past year or not | 0 | 1 | 0.3 |
| read_print | whether one reads printed books for the past year or not | 0 | 1 | 0.9 |

The variable "age" is chosen as people of different ages have different lifestyles and habits. Older people who are retired may have more freetime than people who are worker full time. On the other side, people working in the educational field and students are required to read more for their careers. Therefore it can be on interesting factor to investigate.

The variable *income* is put into groups by the raw data, which is kept for the cleaned data. The variable *employment* is also worth analyzing, as both are factors of one's working status. Since it is anticipated that people with higher incomes are the ones with more knowledge, but those that work more often do not have as much free time to read as those retired and those in labor force. Same for the variable *education*. People at higher educational level builds a stronger habit of reading, which is worth checking. The variable has three groups: *Under high school*, *high school graduate* and *college graduate*.

The variables *read_audio*, *read_ebook* and *read_print* are dummy variables for the model, only taking on the value of 0 and 1. With a 0 indicating that people do not read the types of the book for the past 12 months, and 1 for they have read the types of book for the past 12 months. As technology develops, people can freely read on their phones or even listen to the books. However, some people like to feel the paper texture and feel rewarded after reading a printed book.

Lastly, the response variable $y$ is the variable *num_books*, which is the data for how many books one has read for the past 12 months. More amount of books can shown one's stronger reading habit.

The categorical variables *employment*,*education*,*income*,*read_audio*,*read_print*,*read_ebook* are treated as dummy variables in the model, which means taking on the value of 1 or 0 depending on whether the individual belongs to the group or not. For example, the variable *education* after cleaning has three levels: *Under High school*, *High school graudate*, and *College graduate*. If the individual has an educational background of *High school graduate*, the variable

$$x_{education}$$

will take on the value of 1.

With over 2800 observations for the data, the statistics have a strong statistical power. However, it is inevitable to have *refused* or *Don't know* as the response, which we are considerign as the *na* cases here. The missing values are removed in the analysis as they provide no meaningful information. After removing the *na* cases in the dataset, there are still over 2000 observations.

## III. Model structure

As the purpose of the study is to analyze one's reading habit, a multiple linear regression model is fitting to the subject. The alternative models include a logistic model. However, for a logistic model, the variable is

set to 0 or 1 and is interpreted as one's probability of reading books. It can be confusing and whether one reads or not cannot show how much one reads, therefore the multiple linear regression model is more fitting.

The model is adopted to the chosen response and predictor variable is written as

$$num_{books} = \hat{\beta}_0 + \hat{\beta}_1 * x_{age} \hat{\beta}_2 * x_{education} + \hat{\beta}_3 * x_{income} + \hat{\beta}_4 * x_{employment} + \hat{\beta}_5 * x_{readaudio} + \hat{\beta}_6 * x_{readprint} + \hat{\beta}_7 * x_{readebook}$$

The meaning of each coefficient is written as:

$\hat{\beta}_0$:intercept of num_book (number of books read for the past 12 months) when all regressors are 0

$\hat{\beta}_1$:the expected change in number of books read when age is increased by one

$\hat{\beta}_2$:the average difference in number of books read beween different education groups

$\hat{\beta}_3$: the average difference in number of books read between different income groups

$\hat{\beta}_4$: the average difference in number of books read between different employment status group

$\hat{\beta}_5$ and $\hat{\beta}_6$ and $\hat{\beta}_7$: the average difference in number of books read between groups that have read the type of book (coded as 1) and the group that did not read the type of book (coded as 0)

Among all variables, *age* is the only numerical variable. The other variables are kept as categorical. Therefore, their coefficients represent the average change between the group that is omitted and the group entitled.

The model is built using the software R studio with the *lm()* function, which fits a multiple regression model. The website where the dataset is taken did (*Kaggle*) and its original source where the survey is taken both did not specify the survey method or the way of collecting the data, therefore the linear model is the simpliest version of it.

## IV. Result

According to the model output, the fitted model can be written as the following:

$$num_{books} = 3.18960 + 0.06973 x_{age} - 1.41697 x_{education_{Highschoolgraudate}} - 2.20891 x_{Underhighschool} - 1.14278 x_{100,000tounder150,000} + 0$$

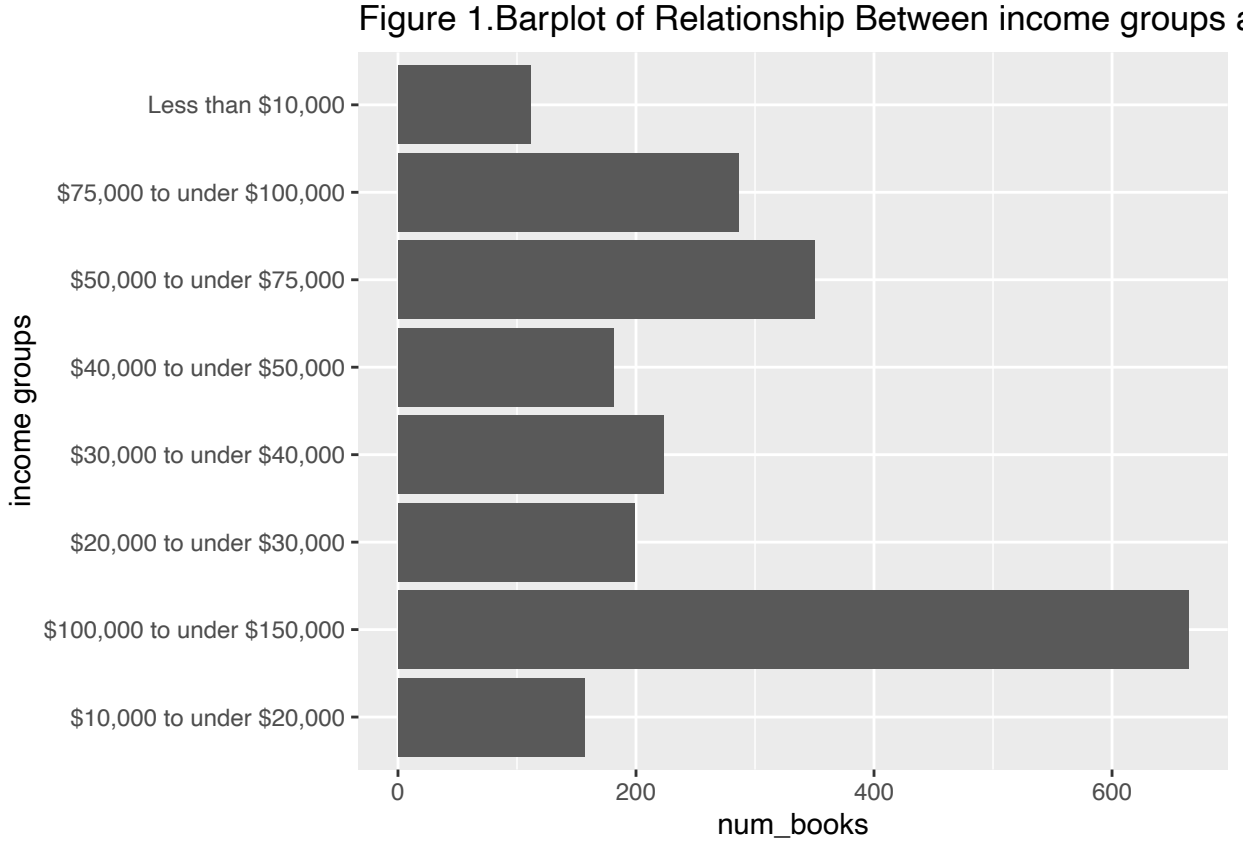
*Table 1* below shows the basic summary information of the logistic regression model. It contains the name of each coefficient, the estimate value, and the p_value.

| Coefficient | coef_num | p_value |
|---|---|---|
| Intercept | 3.18960 | 0.3430280 |
| age | 0.06973 | 0.0543030 |
| education:High school graduate | -1.41697 | 0.2251730 |
| education: Under high school | -2.20891 | 0.2799010 |
| income: $100,000 to under $150,000 | -0.14276 | 0.9478910 |
| income: $20,000 to under $30,000 | 0.69778 | 0.7844900 |
| income: $30,000 to under $40,000 | -1.61010 | 0.6419230 |
| income: $40,000 to under $50,000 | -1.76049 | 0.5032070 |
| income: $50,000 to under $75,000 | -0.52454 | 0.8215990 |
| income: $75,000 to under $100,000 | -2.48951 | 0.3102510 |
| income:Less than $10,000 | 0.95954 | 0.7468270 |
| employment: not in labor force | 4.91967 | 0.0014240 |
| employment: unemployed | 1.55475 | 0.3021130 |
| as.factor(read_audio) | 5.35502 | 0.0001270 |
| as.factor(read_print) | 10.14750 | 0.0000006 |
| as.factor(read_ebook) | 8.70703 | 0.0000000 |

*Table 2* below shows the proportion value of different education groups in categorical variable *education*. The bigger the number is, the more the books are read.

| education | num_books_mean |
|---|---|
| College graduate | 21.53050 |
| High school graduate | 18.33044 |
| Under high school | 15.52558 |

*Figure 1* below shows the relationship between *income* and *num_books*. Shown by the graph, the group with the highest income interval reads the most amount of books.




## V. Discussion

### Summary

In this study, we investigate the potential factors that influence one's reading habit. We started by looking at the variables and fitted them into a multiple linear regression model. We then moved on the describe the result of the model output by interpreting the estimate coefficients. Unfortunately many of the variables are not statistical significant under a significance level of 0.05 and therefore do not have significant correlation with the response variable number of books one reads for the past one year. More details will be explained in the following part:

## Conclusion

*Output for Logistic Regression Model Table 1* summarizes the name, coefficient estimate and the p-value for each of the variables used in the model. The intercept term represents the number of books one reads at age 0, in the College graduate category, with less than $10,000 income, employed and do not read any type of the books. In this case, the coefficient for intercept has no practical meaning. The person with above circumstance do not exist and therefore it is meaningless to interpret the intercept coefficient.

The variable *age* has estimate coefficient of 0.06973, which means as one's age increase by one, the average number of books one reads increase by 0.06973. However, the p-value for this coefficient is 0.343028, which by a significance level of 5%, the null hypothesis that the coefficient equals to 0 fails to be rejected. But with a higher significance leve, the null hypothesis can be rejected, indicating significance relationship between the number of books read in a year and one's age.

For the variable *education*, which is a categorical variable, the group *College graduate* is omitted in the table to show comparison between the groups. The estimate coefficents have shown that with a *High school graduate* education and a *Under high school* education, the number of books one reads in a year decreases by 1.42 and 2.21 on average. This makes sense with reality as people in higher educaitonal field needs to read more in order to widen their knowledge. The estimate coefficients for different *income* groups have different coefficients. Most groups have a negative coefficient, but people with *income $20,000 to under $30,000* and *income Less than $10,000* have positive coefficient.

The coefficient estimate for the variable *employment* both have positive coefficient, which can be interpreted as compared with people *employed*, people *unemployed* and *not in labor force* reads more books on average. This is understandable as people *unemployed* have more free time, and those that are retired and are students, which are counted as *not in labor force* are required to read more and have more time to read.

Lastly the variables *read_audio*, *read_print* and *read_ebook* are dummy variables, taking on the value of 0 and 1. Their coefficients are interpreted as the average difference in number of books read between groups that have read the type of book (coded as 1) and the group that did not read the type of book (coded as 0). It appears that *read_print* has the highest postive coefficient estimate compared to *read_audio* and *read_ebook*. This shows that people do read print books more, then ebook, then audio books. All these three variables are statistically significant with small p-values.

Even most variables have rather large p-values and are statistically insignificant, the estimate coefficients give an idea on what are correlated with the number of books one reads in a year. The large p-values of many variables show a poor variable selection, however, out of all variables given in the raw dataset, the variables chosen are the ones that potentially have a correlation with the response variables.

Unfortunately many variables are statistically insignificant, however this does strengthen the believe that regarding one's position in the world, reading has always been a significant and helpful hobby for people to gain knowledge from. A trait of someone does not define one's reading habit and therefore does not define one's knowledge. However, in extension, this study helps show people's reading preference. People still tends to read priinted books more than other types of books. According to the website where the data is originally taken from, "in total, 34% of Americans have either read an e-book or listened to an audio book in the last year, but relatively few Americans read books in these digital formats to the exclusion of print books", and among all American adults, "80% read for pleasure (35% nearly every day)" (Perrin). Reading is not a forced habit or hobby one should have, it is an entertainment method for people to double gain knowledge from. Enjoy reading, regarding what you are.

*Education and number of books read in each category*

*Table 2* shows the average number of books read in each educational group. It matches with the prediction that people at higher educational group reads more books in a year. The change of 3 books per year is not that significant itself, however, with other identities and factors, one can read a lot more.

*income and number of books read in each category*

The barplot, *Figure 1* studies how different income groups relate to the number of books one reads in a year. The largest amount of books read was the group with highest income *$100,000 to under $150,000*. This

suggests that people with higher income reads more books on average, which can be interpreted as people with higher amount of book reading gains more money.

## VI. Weakness and Next steps

This analysis has a lot of weakness and limitations. The large p-values of many variables show a poor variable selection, where the variables do not have a significant correlation with one's reading habit. To furthur develop the model, more variables should be incorporated into the data and reconsidering selecting the variables. The reason this step was not done in this analysis was due to the lack of variable selection range given by the raw dataset. To make better models, other surveys can be considered.

Also as the dataset was taken from other websites with no direct link to the survey or an introduction on how the data was collected, the model can only be build as the simpliest multiple linear model with no finite population or other procedures. In order to further complete the model, penalties like AIC or BIC can be considered, or cross validation, which helps give an idea on the model's predicting ability.

## VII. References


Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

L. (n.d.). Gorky once said: Books are the stepping stones to human progress. Retrieved December 09, 2020, from https://lang-8.com/1243601/journals/329105475969855823617592253472811287598

Perrin, A. (2020, May 30). Majority of Americans Are Still Reading Print Books. Retrieved December 20, 2020, from https://www.pewresearch.org/internet/2016/09/01/book-reading-2016/

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.

Vipul. (2020, September 03). Reading habit Dataset. Retrieved December 09, 2020, from https://www.kaggle.com/vipulgote4/reading-habit-dataset

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686