Analysing Factors Correlated to One's Reading Habit

Ke Deng 1004957488

21/12/2020

Abstract

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

Keywords

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

Important links:

Github link: https://github.com/kristiadeng/Reading-Habit Data link: https://www.kaggle.com/vipulgote4/reading-habit-dataset Relevant study: https://www.pewresearch.org/internet/2016/09/01/book-reading-2016/

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: https://github.com/kristiadeng/Reading-Habit

Methodology

II. Data analysis

II.A Introduce Dataset

The dataset chosen for the analysis is from the link below: https://www.kaggle.com/vipulgote4/reading-habit-dataset

The data was originally taken from the website: https://www.pewresearch.org/internet/

However, there is no further detail on how the data was collected and the target, survey, or other population details. By one potentially relevant article from the original website *One-in-five Americans now listen to audiobooks* by Andrew Perrin, which the dataset is possible to be used on, the survey was taken in the United States. The study was done focusing on people's preference in reading, and the resulting conclusion includes the population of college graduates that listen to audiobooks increases, the income group with more than $750,000 per year reads more audio books, the male population that listens to audio books increases. The final result shows that audiobooks are becoming a trend in the United States (Perrin). The target population is the United States population, and the sample was those that responded to the survey.

Another report done by Andrew Perrin is *Book Reading 2016*. The report also discusses Americans' reading preferences. The study shows that people are starting to use more ebook or audiobook platforms, but the majority amount of people remain reading print books (Perrin,1).

Both studies can help connect with the current report, but the focuses were different as this report will be focusing on factors potentially correlated with the number of books one reads in a year, while the two other reports focus on reading preferences.

II.B Strength and Weakness

By choosing this dataset, the strength is that there are over 2000 observations, which is large enough to show the significance of the coefficient estimates. The study was also able to show society's current situation and how people of different groups have different reading habits. By the subject of the study, people are more willing to answer this type of question as most people read more than 1 book within a year, therefore there are many volunteers, adding the number of observations.

The inevitable weakness is the lack of information. There is no clear indication of how the data was collected and the possible survey questions. Also, even the two relevant studies show that the target population was the United States population, it cannot be assumed and use finite population to strengthen the model. However, to draw a correlation with the factors, 2000 observations is a rather considerable size of the study.

II.C Variable selection

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 renamed in the cleaning data process as *age*,*education*,*income*,*employment*,*numbooks*,*audio*,*ebook*,*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

variable_name	meaning	min	max	mean
employment	one's employment status	0	0	0.0
numbooks	the number of books individual reads for the past year	1	97	19.4
audio	whether one listens to audio books for the past year or not	0	1	0.2
ebook	whether one reads ebooks for the past year or not	0	1	0.3
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 free time than people who are working 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 an interesting factor to investigate. The study uses the variable *age* instead of *age_group* as a numerical variable is able to show wheter as people gets older, they tend to read more books, rather than people in specific age group reads more books. However, the variable *age_group* is appropriate to use in the model also.

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 the labor force. There are three groups after cleaning for the variable *employment* : *employed*, *unemployed* and *not 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 *audio*, *ebook*, and *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 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 *numbooks*, which is the data for how many books one has read for the past 12 months. The higher amount of books one has read for the past year represents one's stronger reading habit.

The categorical variables *employment*,*education*,*income*,*audio*,*print*,*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 graduate*, 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 considering 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, which is able to draw conclusions like "Law of Large number" or "Central Limit Theorem".

A plot of the cleaned

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. Also, with a logistic model set up, the binary response variable has value of 0 or 1, where the 0 means people has not read a book for the past year, and 1 for has read a book for the past year. Based on the variable we have in the dataset, there is no "0" value of the amount of books one has read in a year, therefore the logistic model do not apply. Also the result will be the probability of one reading a book in a year, which is rather meaningless. Therefore the multiple linear regression model is more fitting.

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

$$numbooks = \beta_0 + \beta_1 * x_{age} \beta_2 * x_{education} + \beta_3 * x_{income} + \beta_4 * x_{employment} + \beta_5 * x_{readaudio} + \beta_6 * x_{readprint} + \beta_7 * x_{readebook}$$

The interpretation of each coefficient (holding other variables constant) is written as:

$\hat{\beta_0}$:intercept of numbook (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:

$$\hat{numbooks} = 3.18960 + 0.06973 x_{age} - 1.41697 x_{education_{Highschoolgraudate}} - 2.20891 x_{Underhighschool}$$

$$-1.14278 x_{100,000 tounder 150,000} + 0.69778 x_{20,000 tounder 30,000} - 1.16101 x_{30,000 tounder 40,000} - 1.76049 x_{50,000 tounder 75,000}$$

$$-2.48951 x_{75,000 tounder 100,000} + 4.91967 x_{notinlaborforce} + 1.55475 x_{unemployed} + 5.35502 x_{audio} + 10.14750 x_{print} + 8.70703 x_{ebook}$$

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

Coefficient	coefficient	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

Coefficient	coefficient	p_value
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(audio)	5.35502	0.0001270
as.factor(print)	10.14750	0.0000006
as.factor(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. By the table below, the *College graduate* group has the largest mean value of number of books read in a year.

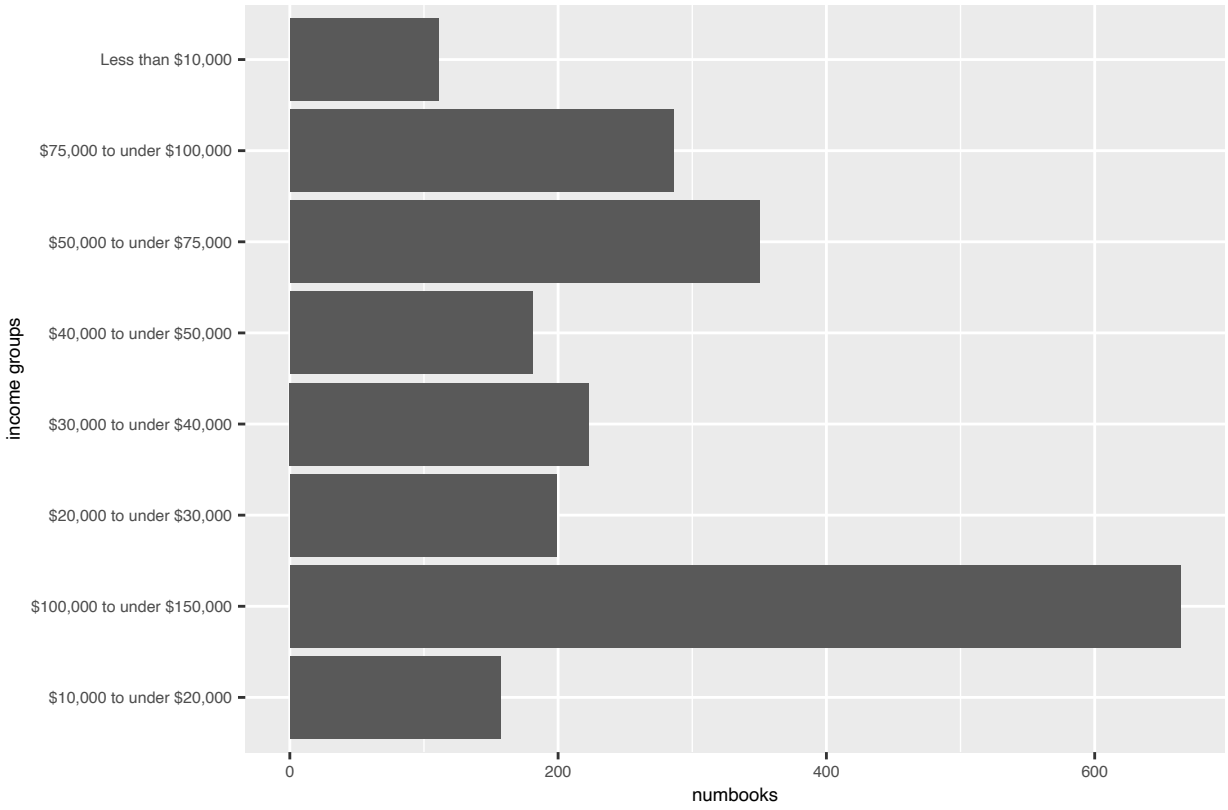
education	mean
College graduate	21.53050
High school graduate	18.33044
Under high school	15.52558

Table 3 below shows the proportion value of different employment groups in categorical variable *employment*. The bigger the number is, the more the books are read. By the table below, the *not in labor force* group has the largest mean value of number of books read in a year.

employment	mean
employed	18.13208
not in labor force	23.88358
unemployed	18.12877

Figure 1 below shows the relationship between *income* and *numbooks*. Shown by the graph, the group with the highest income interval reads the most amount of books. The graph shows that the income group of *$100,000 to under $150,000* reads the most amount of books in a year.

Figure 1.Barplot of Relationship Between income groups and number of books read in a year in total



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 Multiple 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 level, 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 *audio*, *print* and *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 *print* has the highest postive coefficient estimate compared to *audio* and *ebook*. This shows that people do read print books more, than 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 habiit 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 printed 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 who 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. 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.

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.

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/3291054759698558236175922534728112875 98

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/

Perrin, A. (2020, May 30). One-in-five Americans now listen to audiobooks. Retrieved December 21, 2020, from https://www.pewresearch.org/fact-tank/2019/09/25/one-in-five-americans-now-listen-to-audiobooks/

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