Assignment 01

Simple Linear Regression: Description

Michael Levine in a guest blog post on the Huffington Post has suggested that the length of books has declined over time because of "the closing of the American mind…has given way to the collapse of the American attention span (thank texting, Facebook and Twitter)". In this assignment you will be focusing on the relationship between the age of a book (predictor) and its length (outcome).

Please submit your responses to each of the questions below in a printed document. Also, please adhere to the following guidelines for further formatting your assignment:

- All graphics should be resized so that they do not take up more room than necessary and should have an appropriate **caption** and **labels**.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document using Equation Editor, Markdown, or LaTeX.

This assignment is worth 15 points. Each question is worth 1 point unless otherwise noted.

For this assignment, you will use the file *goodreads-2016.csv*. This file contains data from Andy's GoodReads entries. The data consists of 12 variables and 291 observations. The variables are:

- `title`: Book title
- `author`: Primary author of the book
- `my_rating`: Andy's GoodReads rating (on a 5-pt scale)
- `avg_rating`: Average GoodReads rating (on a 5-pt scale)
- `publisher`: Publishing company
- `binding`: Book binding (Harcover or Paperback)
- `pages`: Length of the book (in pages)
- `year_published`: Year the book was published
- `month_read`: Month Andy finished reading the book
- `year_read`: Year Andy finished reading the book
- `bookshelf`: GoodReads bookshelf (to-read; currently-reading; read; quit-reading)

Preparation

Before carrying out any analyses, create a predictor called `age` that indicates the age of the book. To do this, subtract the year that the book was published from the current year (2016). This variable (not `year_published`) should be used in all analyses for this assignment.

Also, filter the dataset (see Assignment 00) so that only the books on the `read` bookshelf are being used for the analysis. After filtering, there should be 225 books in the dataset.

---

1. Create a plot of the marginal distribution of book length. Make sure your plot has a caption.
2. Examine the structure and formatting of Table 1 in the article: Snedker, K. A., Herting, J. R., & Watson, E. (2009). Contextual effects and adolescent substance use: Exploring the role of neighborhoods. *Social Science Quarterly, 90*(5), 1272—1296. Notice that variables are presented in rows and summary statistics are presented in columns. Mimic the format and structure of this table to create a table to present the numerical summary information for the marginal distributions of book length and age. Provide the same measures for these variables as is given in Table 1. Re-create the formatting of Table 1 as closely as you can. Finally, make sure the table you create also has an appropriate caption. **(2pts.)**
3. Create a plot of the distribution of book length *conditioned on* age (i.e., a scatterplot). Make sure your plot has a caption.
4. Describe the relationship between age and book length. Be sure to comment on the structural form, direction and strength of the relationship. Also comment on any potential observations that deviate from following this relationship (unusual observations or clusters of observations). **(2pts.)**
5. Compute and report the Pearson correlation coefficient between age and book length.
6. Based on your answer to the Question #4, is the Pearson correlation coefficient an appropriate summary measure of the relationship? Explain. (Hint: Pay attention to the structural form!)
7. Regresss book length on age. Write the *fitted equation* using Equation Editor (or some other program that correctly types mathematical expressions).
8. Interpret the value of the intercept from the regression equation using the context of the data.
9. Interpret the value of the slope from the regression equation using the context of the data.
10. Compute and report the value for R^2 based on values from the ANOVA decomposition.
11. Interpret the value of R^2 using the context of the data.
12. Craft a short reply to the editor of The Huffington Post regarding whether this evidence supports Mr. Levine's thesis that the length of books has declined over time. This reply should be no more than a paragraph and should cite results from the regression anlysis to support your claims. (Hint: Pay attention to the slope.) It should also specifically address the validity of Mr. Levine's thesis. **(2pts.)**