Assignment 02

Simple Linear Regression: Inference

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 14 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.

Finally, fit a regression model using age to predict book length.

Model-Level Inference

1. Using symbols, write the null hypothesis that is tested by the F-statistic in this analysis.

2. Write no more than three sentences (to be included in a publication) that summarizes the results of the omnibus analysis. A summarization of the results includes a written description of what is being tested by the F-test and the statistical results. At a minimum report the F-statistic, *df*, and p-value. A summary should also indicate what the statistical results suggest about the tenability of the null hypothesis and what this means about the potential relationship between age and book length. **(2pts.)**

Coefficient-Level Inference

3. Using symbols, write the null hypothesis that is tested by the t-statistic for the intercept.
4. Based on the results of the t-test, what do the data suggest about the tenability of the null hypothesis for the intercept? Explain.
5. Using symbols, write the null hypothesis that is tested by the t-statistic for the slope.
6. Based on the results of the t-test, what do the data suggest about the tenability of the null hypothesis for the slope? Explain.
7. By referring to the study design, explain why the inferences about the relationship between age and book length are likely biased and do not apply to books generally.

Analysis #2

For this analysis, you should fit the same model, regressing book length on age, to all 291 books in the original dataset (no filtering).

8. Based on the results of the t-test, what do the data suggest about the tenability of the null hypothesis for the slope in this analysis? Explain.
9. Given the larger sample size in this analysis, would the inferences be *less* biased from this analysis? Explain.

Displaying Results

10. Examine the structure and formatting of Table 1 in the article: Garcia, D. R., McIlroy, L., & Barber, R. T. (2008). Starting behind: A comparative analysis of the academic standing of students entering charter schools. *Social Science Quarterly, 89*(1), 199—216.

Notice that models are presented in columns (in the article 6 models are presented). Predictors used in the models are presented in rows, and so are the model-level summaries (e.g., N, F, R^2). Also note that the intercept (‘Constant’) is the last term presented in the table, generally because it is the least important coefficient. Blank cells indicate that the model does not include that particular predictor.

Mimic the format and structure of this table to create a table to present the numerical information for the two models you fitted in this assignment. Re-create the formatting of Table 1 as closely as you can. Instead of giving the adjusted R^2 value, provide the unadjusted R^2 value. Make sure the table you create also has an appropriate caption. **(2pts.)**

11. Create a plot that displays the regression line from both analyses. This plot should also include the data ($n = 291$) plotted as a scatterplot. The data should be semi-transparent, and both regression lines should be completely opaque (non-transparent). The lines should have different linetypes or use different colors so that they can easily be differentiated. Give your plot an appropriate caption. **(2pts.)**