Measuring Paper Discoloration Over Time

May 30, 2014

1 Question

Is there a linear correlation between the year that a book was printed and how the color of its paper has changed over the years?

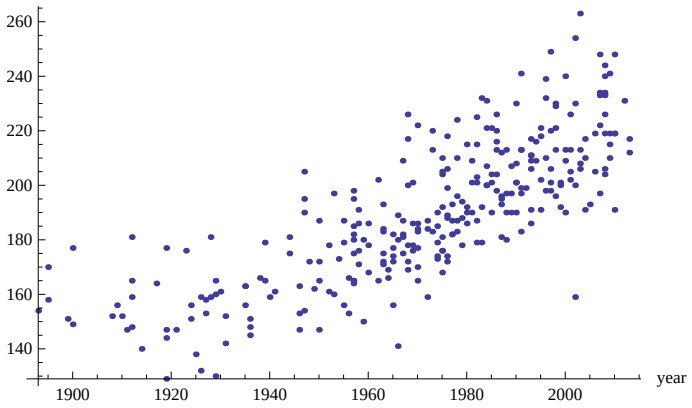
2 Experimental Design

We took a simple random sample of 300 values from 0 to 999.99 with two decimal places to select the books from the Oliver Wendell Holmes library. We used a random number generator online to generate these values. Not every number matched with a book in the library, so we would always take the number randomly selected and choose the next highest book in the library. For example, our simple random sample gave us the value of 29.22, but the actual book that we used was 29.7 C6 because that was the next book after 29.22 in our library. Additionally, if the book had glossy or laminated paper we would skip it and choose the book with the next highest call number. By using this method along with a large sample size, we were able to ensure that our sample would represent the range of ages of books that our library has fairly accurately, without us having to manually find older books, thereby potentially introducing a source of bias. Once we selected our book, we recorded the actual call number and the year it was printed, or the most recent edition published if the printed date was not given. Once we recorded all the information, we opened to the title page or another page in the front that was mostly empty space without anything printed on it, measure the color value of the paper at that point, and move on to the next book.

In order to measure the color of paper objectively and quantitatively, we used a Lego Mindstorms NXT robotics kit with a HiTechnic Color Sensor V2. Once every ten seconds, the sensor would detect the reflected light from a surface placed in front of it. In order to ensure that our results would remain consistent, the height of the sensor and the angle of the sensor were kept constant. (The consistent angle also ensured that the sensor would not pick up light being reflected directly back at it.) Differences in light levels from external sources were also kept to a minimum, but as the sensor uses an LED much brighter than the external light, the external light level should not influence the sensor reading by more than a few units anyway, and what little testing we did on this topic seemed to support that claim. The higher the value, the more light was reflected, and the whiter the paper is.

3 Graphs

When plotted on a graph with publishing year on the x-axis and the sensor value on the y-axis, the data appears as the following.



We can calculate the linear regression equation for this data using

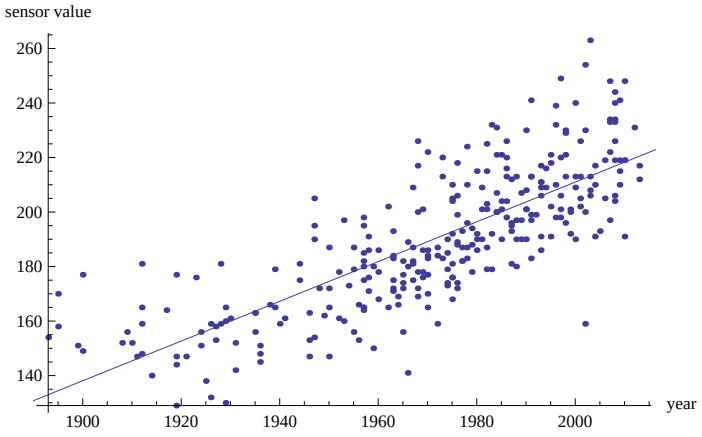
$$y = \beta_1 x + \beta_0$$

where y is the sensor value and x is the year. The formula to find β_1 and β_0 are, respectively,

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{(300)(112476998) - (591335)(56978)}{(300)(1165819005) - (112476998)^2} \approx 0.72886$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 189.9 - (0.72886)(1971.11) \approx -1246.76$$

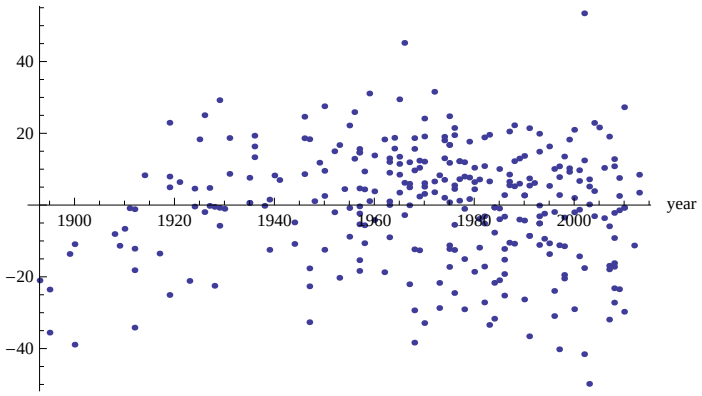
This equation seems to fit our data reasonably well. The line given by the equation, along with the R^2 value is shown below.



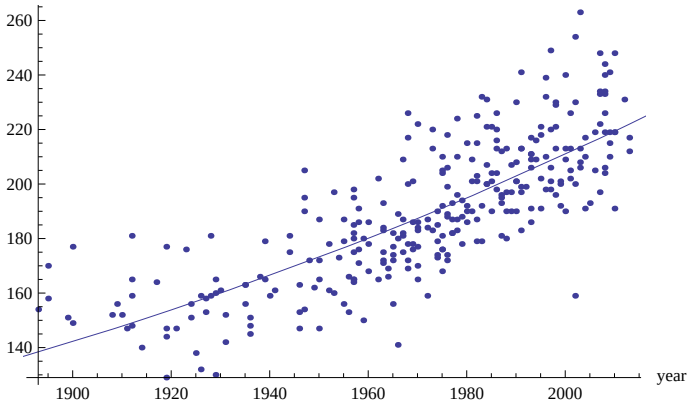
$$y = 0.72886x - 1246.76$$

$$R^2 = .6095234$$

From this, we can calculate residuals for each point on the graph. Furthermore, by plotting the residuals at the value they correspond to, we can determine if patterns emerge, which would indicate that the linear model is an incorrect model.



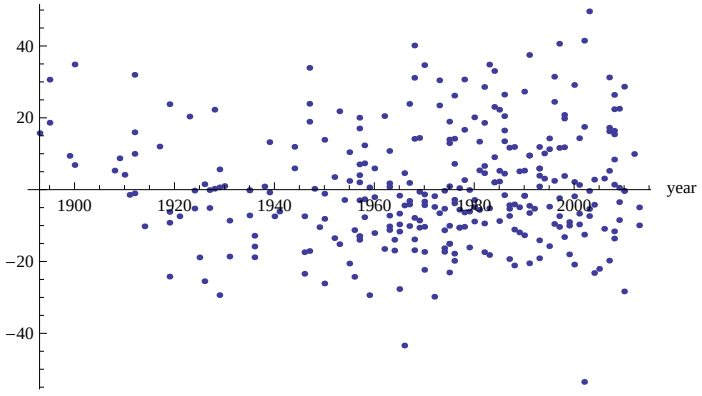
There does not appear to be a noticeable pattern in the data.
The only non-linear model which we thought might have fit the data better was the exponential model, shown below.



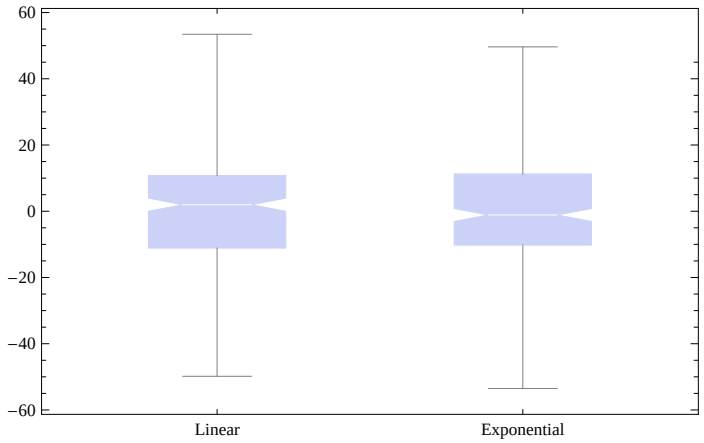
$$y = 0.0794e^{0.00394x}$$

$$R^2 = 0.6272006$$

As for the residuals, no significant pattern was revealed for this model either.



The box plots for both sets of residuals seemed fairly close to normal. The plots for both are shown below.



4 Analysis

Since we are studying the relationship between discoloration and age, our parameter of interest is β, which we define as the slope between the age of the book and the sensor value. Consequently, our null hypothesis is that there is no true linear relationship between age and discoloration. Our alternative hypothesis is that there is some positive linear correlation between the two variables. In other words,

$$H_0 : \beta = 0$$
$$H_a : \beta > 0$$

We also had to ensure that our collected data met the conditions for being able to run this test accurately.

- Observations are independent because it is an SRS of 300 books.
- Relationship between the age of books and the corresponding sensor reading appears to be linear as evidenced by the lack of a pattern in the residual plot.
- Consistent spread in residuals in our residual plot shows that the variability of the errors is fairly consistent.
- A boxplot of the residuals is fairly symmetric, suggesting that even accounting for outliers, it is not unreasonable to assume prices vary normally about LSR line.

Next, it is necessary to determine the t-value that we can use for calculating the probability that there exists a linear correlation in our data. Here, b is used to represent the test statistic of our sample (or, in other words, the slope of the least squares regression line).

$$t_{n-2} = \frac{b - \beta_0}{SE_b} = \frac{b}{SE_b}$$

Since we know the value of b, we only need to calculate SE_b. x and y represent year and sensor value respectively.

$$SE_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0.033790$$

If we plug these values back into the formula for t, we get:

$$t_{n-2} = \frac{b}{SE_b} = 21.57$$

The last piece of information we need is the degrees of freedom, which are simply calculated by:

$$df = n - 2 = 298$$

For a t-distribution of 298 degrees of freedom and a t-score of 21.57, we get a p-value of ≈ 0.

An exponential model also matched the data fairly well. Whereas the linear regression explains only 60.95% of the variation in whiteness, an exponential model would explain 62.72% of the variation in whiteness. The residual plot also shows a slight pattern, consistent with an exponential model. Although an exponential regression would probably have been a slightly better choice, the difference in explanatory power is so small compared to the difficulty of running a significance test that we decided to pursue only the linear model.

5 Conclusions

Assuming the null hypothesis is true, there is essentially no chance of getting a linear association this strong or stronger. Compared to an α value of .05, our p-value of 0.000 is low, allowing us to reject the null hypothesis and accept that there is a true, positive linear relationship between the age of the book in years and the whiteness of the paper. Although the r^2 value is only .60952, the amount of data and the clear correlation between age and whiteness allow us to conclude that there is a linear relationship.

Interestingly, the linearity of the data increased over time. When the correlation between age and whiteness is plotted by decade, there is a clear upward trend. A least-squares regression line (excluding the outlier at 1930, .510) on decade explains 40.58% of the variation in correlation. This result implies that the strength of the linear relationship increases as the age of the books decreases. This is probably the reason why the exponential model is slightly more accurate than the linear model. In addition, this relationship speaks to the faults of this study: paper quality and printing techniques have probably changed over time. Biases probably also come from the fact that very yellowed or decayed books are eliminated from the collection. While the data accurately describe overall yellowing in the Phillips Academy library, they probably do not describe the specific pattern of yellowing followed by any one book. In other words, we can draw conclusions about the state of the OWH library, but we cannot make predictions about the changes in any specific book as time goes on.

6 Reflection

Our project, for the most part, was conducted and run properly. Initially, we ran into issues because the place on the book where we measured results in different sensor readings. Near the binding in darker conditions, the sensor value is much higher because paper there is less exposed to factors that would affect discoloration. On the outside of the book under normal lighting conditions, the results were consistent. This discrepancy caused us to rerun a few of our samples because the values recorded initially were too high when compared to the values taken from those obtained on the edge of the book.

Furthermore, the type of paper added a possible lurking variable. Because we have no knowledge of the whiteness of each book when it was printed, we are making all of our assumptions on deterioration based on the idea that all books printed since 1890 have been printed on the same type of paper that was perfectly white at the time of printing, or that any differences in paper type do not affect the rate at which paper changes color.

Another potential lurking variable would be how often the book in question was used. In some extreme cases, we noticed that exceptionally non-deteriorated old books were rarely, if ever, checked out of the library.

Additionally, we will not be able to extrapolate beyond approximately 1890. We only collected data from books printed between 1893 and 2013, so trying to estimate the white value of a book printed in 1200 would not result in an accurate approximation. Extrapolation also is not possible because the y-intercept is -1246.7. That means that a book from the year 0 CE would have an extremely negative white value, which is not possible. For a book from 1890 through 2013, the linear model offers a fairly good fit. The r^2 value tells us that .60592 of the variation in white value is explained by the year the book was printed.