UNIVERSITY OF WATERLOO

FACULTY OF ENGINEERING



DEPARTMENT OF MECHANICAL AND MECHATRONICS ENGINEERING

**ME 202 Project Report**

Prepared By:

Kamin Andres Cheung Zhang

Kristoff Stefon Sahadeo

Victor Sung

**Introduction:**

According to an article titled “The cost of a Canadian university education in six charts” written by Mark Brown, university of students spend around $773 per year on books (Brown). Similarly, another article published by the Ottawa Citizen has stated that “a typical university student could spend up to a shocking $1000 per semester on textbooks alone” (Yan).

In this project, a statistical study was performed to determine whether there is a relationship between the number of pages in a college textbook and its price. As previously stated, university students spend a significant amount of money on books alone per semester, which they might only use for a few times. As a result, there was an interest to determine key factors contributing to the high cost of university textbooks. The results from the study will demonstrate whether the cost of materials in books truly has a significant impact on the listing price of a particular book. Since the study was conducted by engineering students, there was a major emphasis on STEM college textbooks published after the year 2000 because current engineering students will mostly buy STEM college textbooks published in recent years for their courses. Therefore, the sample for this study was obtained from the Davis Centre Library because it houses resources for engineering, mathematics, and science at the University of Waterloo. This ensured the sample represented all college STEM textbooks.

**Methods:**

There were two major variates used in this study. They are as follows:

***p –* page count for each STEM college textbook published after the year 2000**

***q* – price (in CAD) of each college STEM college textbook published after the year 2000**

The variables above are for the entire population (all STEM college textbooks) and are therefore population variates, however it is too time consuming to collect all values of p and q, therefore, a representative sample was taken from this population. The variates were as follows:

***x –* page count for each book in the sample (explanatory variate)**

***y –* price (in CAD) of each book in the sample (response variate)**

In this study, linear regression was used to determine the relationship between *x* and *y.* The formula to determine this relationship required the mean values for both *x* and *y*. These were denoted as follows:

**– mean page count for each book in the sample**

**– mean price (in CAD) for each book in the sample**

(where are the equivalent population parameters for the means)

The sample was obtained by carrying out stratified sampling on the main floor of DC Library. DC Library was chosen since it only contained college textbooks related to Science, Technology, Engineering and Mathematics (STEM), thus making data collection easier for the researchers since they did not need to filter STEM books from all other genres. Each stratum consisted of each row of shelves and 2 random samples were taken from each. It was ensured that the book was not published before the year 2000 during data collection. This was due to the fact the current Engineering students typically purchase newer editions of textbooks, rather than older ones, thus making the data more relevant to the research problem. It was also easier to find prices online for newer text books.

To carry out random sampling, a system was developed for each shelf so that a number was assigned to every book in each row in ascending order starting from the top left to the bottom right (relative to the researcher). The numbers were only assigned to books that were published from the year 2000 and above. An online random number generator, powered by Google (“Random Number”) was then used to randomly select two (2) books from this stratum by first inputting the total number of books assigned per shelf as the maximum number, then randomly generating two numbers. The books corresponding to these numbers were then sampled.

Each book was sampled by first recording the full title of the book and then recording the page number, *x*, on the last page of the book. The book was then found on Amazon’s website (“Amazon.com: Online Shopping”) and the lowest price, *y*, for a brand-new version of the sampled book was then recorded. This data was tabulated and used for analysis afterwards. When determining the price, only softcover versions of the books were considered to ensure that this did not have an influence in our data (since hardcover books are typically more expensive).

**Results:**

**Data Preparation:**

Additional metadata was collected during the data collection process. Since only the page count and lowest price data are used for regression, these two variables were moved to a new dataframe. The book title was also included in the dataframe for identification purposes. The statistics for the 2 regression variables can be seen in Table 1.

Table 1: Regression Variable Statistics

A screenshot of a cell phone

Description automatically generated

The Pearson correlation coefficient of the 2 variables was calculated using “.corr”(a function from the pandas library). The correlation coefficient obtained is **0.312**.

**Plotting:**

The data was plotted in a scatter plot.

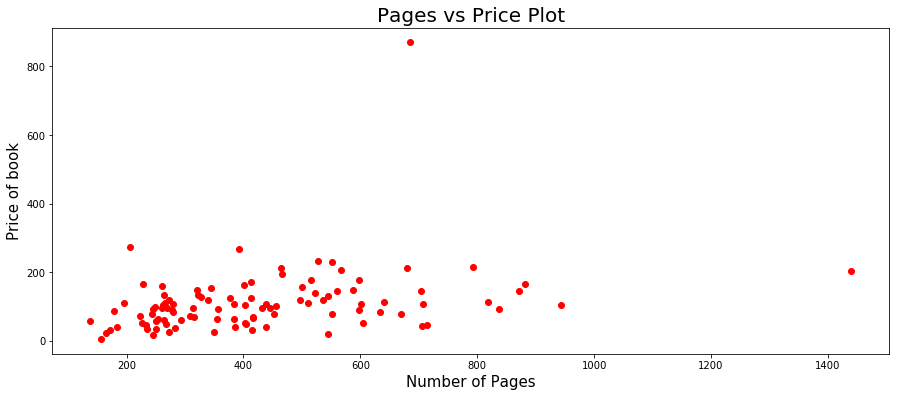


Figure 1: Pages vs Price plot

Outliers were identified when comparing relationships between the 2 variables (Renze). A scatter plot helped to identify any outliers. There were 2 outliers as shown in Figure 1.

**Fitting Regression Line:**

A regression line was fitted on the dataset. The line of best fit was found with the normal equation. The normal equation is:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

The regression parameters **β0**and **β1**were obtained were **β0**is the y intercept of the regression line and **β1**is the slope of the regression line.

**β0**= 51.22412279

**β1**= 0.1407804

The plotted regression line can be seen in Figure 2.

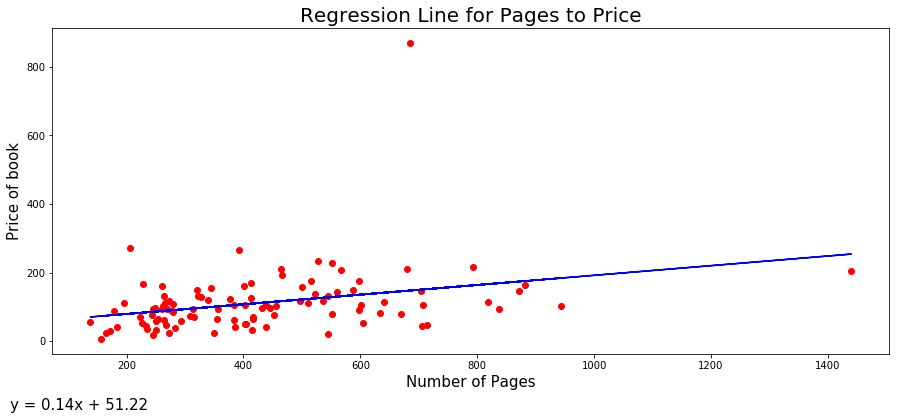


Figure 2: Regression Line for Pages vs Price

**Regression Line with Outliers Removed:**

Outliers can significantly affect the line of best fit for a dataset. The two outliers identified during plotting are removed from the data to explore the effects they have on the regression line.

The new correlation coefficient was recalculated: **0.319**

The dataset with the outliers removed was plotted as seen in Figure 3.

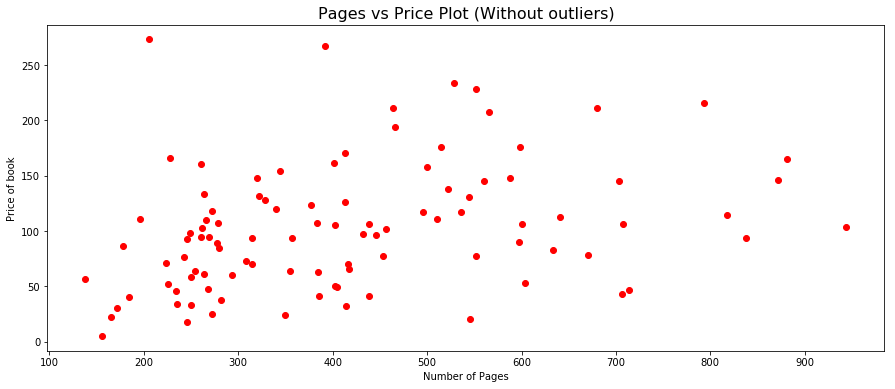


Figure 3: Pages vs Price plot (with outliers removed)

Another regression line was fitted on the dataset with the outliers removed.

The regression parameters **β0**and **β1**were obtained for the new regression line is:

**β0**= 62.161

**β1**= 0.0985

The new regression line was plotted with the original regression line and can be seen in Figure 4.

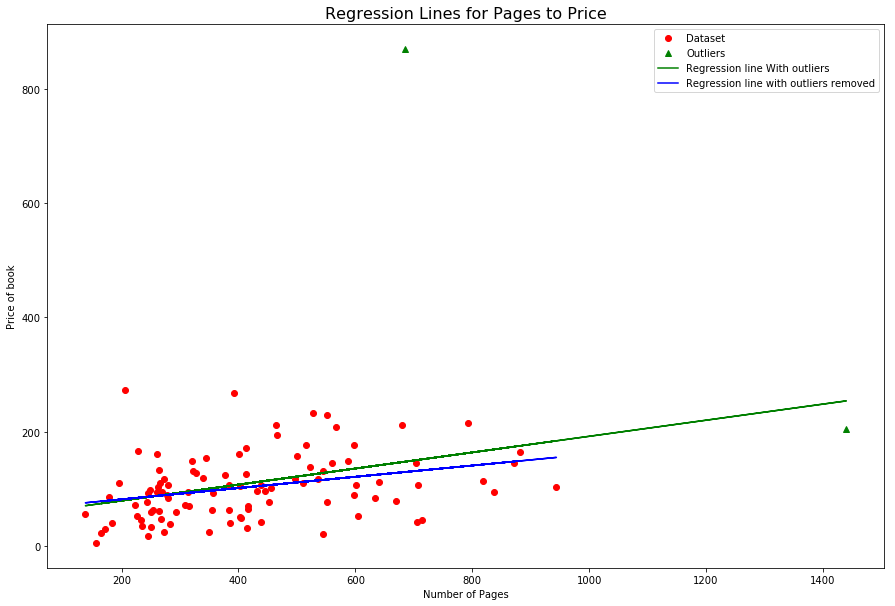


Figure 4: Plot of both regression lines

The standard error of both regression lines was calculated. Standard error is defined as:

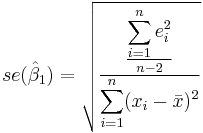


Figure 5. Standard Error Equation (“Simple Linear Regression Analysis”).

Standard Error with all data: **0.001339**

Standard Error with outliers removed: **0.00159**

The outliers did not seem to have a significant effect on the standard error.

**Confidence Interval and Hypothesis Tests:**

A confidence interval was determined for **β1**and a hypothesis test was also conducted on **β1**. The **β1**obtained on the dataset without outliers was used to remove the effect of the outliers that could have on the confidence interval and hypothesis test results.

The 95% confidence interval for β1 was calculated to be **0.0984 ± 0.00316**.

Therefore, the confidence interval is (CA$0.0953, CA$0.1016).

This meant that there was a 95% confidence that an increase in one page is associated with a mean increase of approximately between 1 and 10 cents on the book's price.

A hypothesis test was conducted on **β1** to determine if there is a significant positive association. If there is no relationship between the number of pages in a book and the book's price, β1 will be equal to zero.

To setup the hypothesis test:

: β1 = 0

: β1 ≠ 0

The following results were obtained:

**β1:** 0.0985

**Standard error:** 0.00159

The t value can be calculated as **β1** divided by the standard error.

**t-value:** 61.906

**The degree of freedom:** 98

**T-test value:** 1.0

**P-value:** 0.0

Since the p-value is less than 0.05,

**is rejected**.

Based on the hypothesis test on β1, it can be concluded that there is a significant positive association between the number of pages and price of the book.

**Conclusion:**

In this statistical study, the main goal was to determine the relationship between the number of pages in college textbooks and their prices. However, there was more focus on STEM college textbooks because it was more relevant for the engineering students conducting the study. Therefore, the population of this study was all STEM college textbooks. The sample consisted of STEM college textbooks from the Davis Centre Library at the University of the Waterloo. The number of pages and price of each textbook in the sample were recorded. The relationship between these two variates with a regression model. The correlation coefficient of the model without outliers was 0.319. Since the coefficient value is close to 0, the relationship between the variates is weak. Additionally, the value of the model parameter without outliers was 0.0985, which means an increase in one page in the textbook is associated with an estimated mean increase in the textbook’s price of CA$0.0985. A hypothesis test for revealed that there was a significant positive association between a textbook’s number of pages and prices. Similarly, a confidence interval test showed that there was a 95% confidence that an increase in one page is associated with a mean increase of between CA$0.0953 and CA$0.1016 in the textbook’s price. The results from both tests demonstrates that there is a positive association between a STEM college textbooks’ number of pages and their prices.

Works Cited

"Amazon.Com: Online Shopping For Electronics, Apparel, Computers, Books, Dvds & More". *Amazon.com, Inc.*, 2019, https://www.amazon.com.

Brown, Mark. "The Cost Of A Canadian University Education In Six Charts - Macleans.Ca". *Macleans.Ca*, 2018, https://www.macleans.ca/education/the-cost-of-a-canadian-university-education-in-six-charts/.

"Python Data Analysis Library — Pandas: Python Data Analysis Library". *Pandas.Pydata.Org*, 2019, https://pandas.pydata.org.

"Python For Data Analysis Part 24: Hypothesis Testing And The T-Test". 2015, http://hamelg.blogspot.com/2015/11/python-for-data-analysis-part-24.html. Accessed 8 Apr 2019.

"Random Number - Google Search". *Google.Com*, 2019, https://www.google.com/search?q=random+number.

[Renze, John](http://mathworld.wolfram.com/topics/Renze.html). "Outlier." From [*MathWorld*](http://mathworld.wolfram.com/)--A Wolfram Web Resource, created by [Eric W. Weisstein](http://mathworld.wolfram.com/about/author.html). <http://mathworld.wolfram.com/Outlier.html>

"Simple Linear Regression Analysis - Reliawiki". *Reliawiki.Org*, 2019, http://reliawiki.org/index.php/Simple\_Linear\_Regression\_Analysis#Fitted\_Regression\_Line.

Yan, James. "Yan: The Unfair Economics Of University Textbooks". *Ottawa Citizen*, 2014, <https://ottawacitizen.com/news/national/yan-the-unfair-economics-of-university-textbooks>.

Balka, Peter “Introduction to Simple Linear Regression” ME202 lecture slides

learn.uwaterloo.ca

Appendix A

Please See Separately Attached File for Appendix A

Appendix B

Please See Separately Attached File for Appendix B