

Investigating the linear relationship between
a swimmer's height and their respective lap times

Student Name

Candidate number: 000205-XXX

Math IA

Investigating Graphs

As part of my high school mathematics' curriculum, I have studied graphs, functions, and intercepts. One mathematical theory that has caught my attention was creating and plotting graphs that contain x and y intercepts. The reason I chose the 50 m freestyle is that it is commonly used in everyday life by professional athletes and amateur swimmers alike. The 50m freestyle also offer a benchmark for speed as it is the shortest distance swim for the FINA world championships. Having studied these graphing applications, I wanted to explore this area in relation to the 2012 FINA Swimming Championships.

Statement of the Task:

The task of the project will be through the form of an investigation that will display the results of swimmer heights and their lap times through the form of linear regressions.

This project will be in the form of investigating a particular type of swimming races known as the 50 meter Men's freestyle races that are taken place around the world. I chose this topic as I have an interest in swimming particularly the free style stroke. I will be taking this data from various sport directories, articles, and websites.

In order to complete this project, I will examine the results from the swimmers in the world with the fastest swimming times in the 50 meter freestyle and present these data on a graph in correlation to their individual heights to produce. Once taken by their heights, I will then measure the relationship between a swimmer's height and their respective lap times. The final step would be to evaluate the data chronologically and discuss the direction regression on the graph through the use of a graphical calculator.

During the duration of this experiment, I may come across problems with obtaining the accuracy of swimmer's heights and finding enough valid dates within the past year. There will also be the issue of accuracy of the swimmer's heights from various sources.

Data Collection:

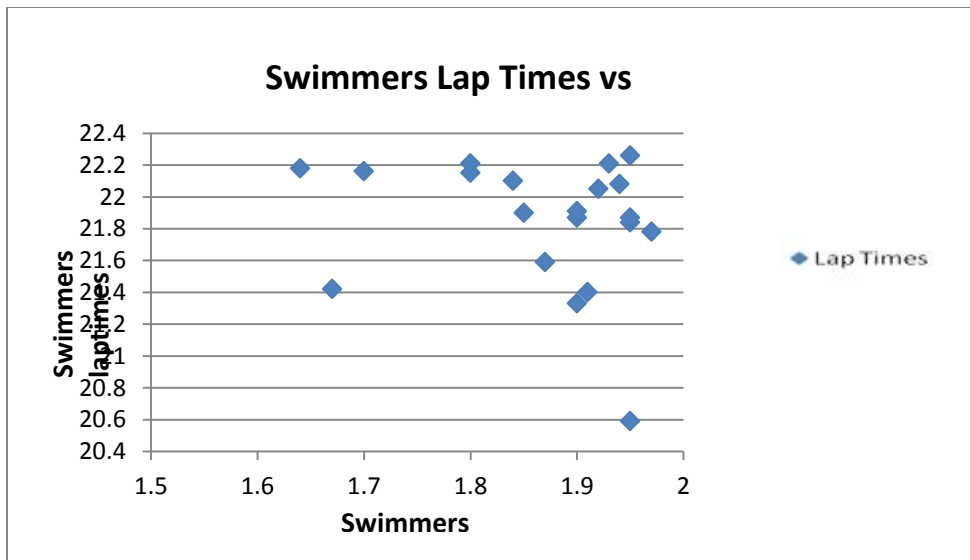
In order for me to conduct this investigation, I needed the swimmer's height and their individual lap times for a specific race known as the 50m Freestyle. The results were taken online from sport biographies, Olympic records, and sport directories.

One such example would be the swimmer "Roland Schoeman" a Russian swimmer who competed in the 50 m Freestyle in the FINA Swimmer's world championships and the 2012 London Olympics. Next I need to find Roland's height; firstly I went to the website FINA.org to find the fastest men swimmers in the world. After searching in their Records Directory, I was able to locate a list of swimmers that have competed in the 2012 FINA Swimming championships with the fastest lap times. Through the London Olympics' Sport directory I was able to search by keyword, country, and name and when I inserted "Schoeman" into their search engine which resulted in a full bio of the swimmer. From the information of Roland's biography I was able to determine that his height was 1.90M or 190 cm. This was the procedure done for a single athlete and the search process was repeated for all of the twenty athletes.

Data Presentations and Analysis:

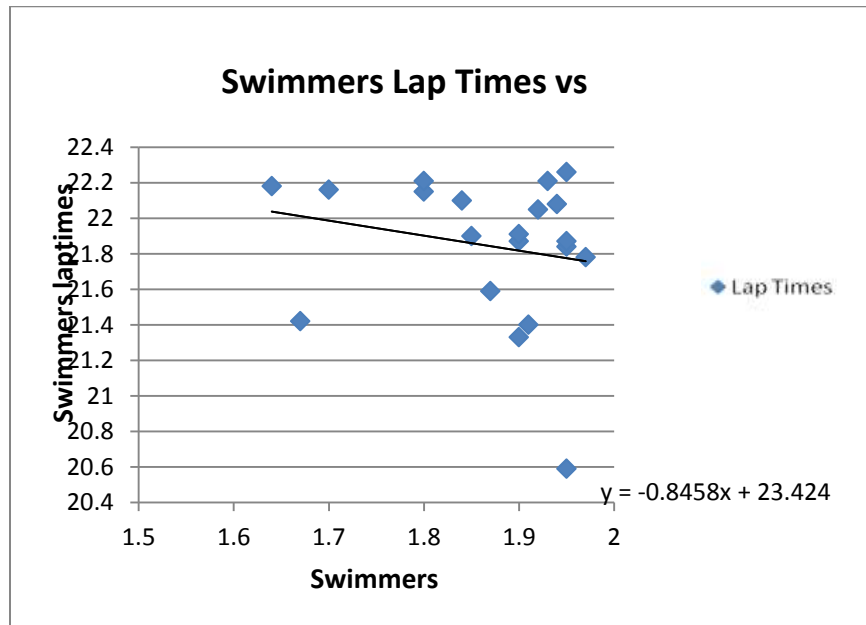
Based on the data listed in the appendices I was able to create a scatter plot graph using Microsoft Excel. Then I was able to manually perform the calculations to check my results. The results are shown below:

Figure 1: Swimmer's Lap times vs. Height.



Based on the above graph, generally a linear pattern occurs with heights, as the greater the height the faster the overall swim time. The next graph display the information with a trend line depicting the swimmer's lap times along with their respective heights. As you can see, a slight negative correlation is formed. Due to the gradient of the regression line the graph is negative. The Linear Regression is calculated through the use of a graphing calculator. The Y axis displays information on the swimmer's respective lap times while the X axis displays the swimmers height. The Linear Regression calculations will be explained below.

Figure 2: Swimmer's Height Trend line Graph



The graph was created through Microsoft Excel to find any discrepancies in my data. I then used Excel to check my manual calculations. Based on the graph above, there is a slightly negative linear relationship between the swimmers height and their respective lap times. However the linear relationship is a very weak correlation as observed by the gradient of -0.8458 . This is a negative relationship and as a result it proves that there is a slim chance of determining that height does have an effect on height. The purpose of calculating the r value was to determine the correlation between the swimmer's height and their lap times. After performing the necessary calculations I observed that there was little to none correlation as the solution of r resulted in 0.20 . The outlier is a data value that is kept away from the group. An example of this would be the point to the top left. If the outlier was removed, the variation will be less. The negative variation will be less so more accuracy in the calculation will be obtained.

This in turn shows the values of r a regression line. Based on the data provided a positive correlation was formed, and the equation $y = -0.0485x + 2.9265$ is established. When calculating the linear equation the one can find the r which is -0.05. Excel gave me the following data and I was able to calculate it manually to verify the values. I decided to calculate the value by hand.

To find the Y formula Use **Stat** then **edit** insert data and **calc linear regression**.

$$r = \frac{s_{xy}}{s_x s_y}, \text{ where } s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}, \quad s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}$$

and s_{xy} is the covariance

The formula above was selected from the IB formula booklet to calculate the linear regression and covariance. In order to use the formula, I need the standard deviation of x (Sx) the standard deviation of y (Sy) and the covariance (Sxy). My calculations for each of these values are shown below.

Data Calculations:

| Height (meters) | Lap time (seconds) | Calculated data: |
|-----------------|--------------------|--|
| x | y | $\bar{x} = 1.95 + 1.90 + 1.91 + 1.67 + 1.81 + 1.97 + 1.95 + 1.90 + 1.95 + 1.85 + 1.90$ $+ 1.92 + 1.94 + 1.84 + 1.80 + 1.70 + 1.64 + 1.80 + 1.93 + 1.95 =$ $\sum \bar{x} / 20$ $\bar{x} = 1.867$ $n = 20$ \bar{y} $= 20.59 + 21.33 + 21.40 + 21.42 + 21.59 + 21.78 + 21.84 + 21.87 + 21.87 + 21.90 + 21.91 + 22.05 + 22.08 + 22.10 + 22.15 + 22.16 + 22.18 + 22.21 + 21.21 + 21.26$ |
| 1.95 | 20.59 | |
| 1.90 | 21.33 | |
| 1.91 | 21.33 | |
| 1.67 | 21.42 | |
| 1.87 | 21.59 | |
| 1.97 | 21.78 | |
| 1.95 | 21.84 | |
| 1.90 | 21.87 | |
| 1.95 | 21.87 | |
| 1.85 | 21.90 | |
| 1.90 | 21.91 | |
| 1.92 | 22.05 | |
| 1.94 | 22.08 | |

| | | |
|------|-------|----------------------------|
| 1.84 | 22.10 | $n=20$ $\bar{y}=21.845$ |
| 1.80 | 22.15 | |
| 1.70 | 22.14 | |
| 1.64 | 22.18 | |
| 1.80 | 22.21 | |
| 1.93 | 22.21 | |
| 1.95 | 22.26 | |

During this calculation, the mean is calculated for each of the values for both the height and the lap times were calculated using the average formula.

| X=Heights | | | Y= Lap times | | |
|-----------------|-------------------|---------------------|------------------|-------------------|---------------------|
| $\bar{x}=1.867$ | | | $\bar{y}=21.845$ | | |
| $n=20$ | | | $n=20$ | | |
| x_i | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ | Y_i | $(y_i - \bar{y})$ | $(y_i - \bar{y})^2$ |
| 1.95 | 0.083 | 0.006889 | 20.59 | 20.59 | 1.575025 |
| 1.90 | 0.033 | 0.001089 | 21.33 | -.515 | 0.265225 |
| 1.91 | 0.043 | 0.001849 | 21.40 | -.445 | 0.198025 |
| 1.67 | -.197 | 0.038809 | 21.42 | -.425 | 0.180625 |
| 1.87 | 0.003 | 0.000009 | 21.59 | -.255 | 0.065025 |
| 1.97 | 0.103 | 0.010609 | 21.78 | -.065 | 0.004225 |
| 1.95 | 0.083 | 0.006889 | 21.84 | -.005 | 0.000025 |
| 1.90 | 0.033 | 0.001089 | 21.87 | 0.025 | 0.000625 |
| 1.95 | 0.083 | 0.006889 | 21.87 | 0.025 | 0.000625 |
| 1.85 | -0.017 | 0.000289 | 21.90 | 0.055 | 0.003025 |
| 1.90 | 0.033 | 0.001089 | 21.91 | 0.065 | 0.004225 |
| 1.92 | 0.053 | 0.002809 | 22.05 | 0.205 | 0.042025 |

| | | | | | |
|---|----------------------------|-------------------------------------|---|----------------------------|-----------------------------------|
| <u>1.94</u> | <u>0.073</u> | <u>0.005329</u> | <u>22.08</u> | <u>0.235</u> | <u>0.055225</u> |
| <u>1.84</u> | <u>-.027</u> | <u>0.000729</u> | <u>22.10</u> | <u>0.255</u> | <u>0.065025</u> |
| <u>1.80</u> | <u>-.067</u> | <u>0.004489</u> | <u>22.15</u> | <u>0.305</u> | <u>0.093025</u> |
| <u>1.70</u> | <u>-.167</u> | <u>0.027889</u> | <u>22.16</u> | <u>.315</u> | <u>0.099225</u> |
| <u>1.64</u> | <u>-.227</u> | <u>0.051529</u> | <u>22.18</u> | <u>.335</u> | <u>0.112225</u> |
| <u>1.80</u> | <u>-.067</u> | <u>0.004489</u> | <u>22.21</u> | <u>.365</u> | <u>0.133225</u> |
| <u>1.93</u> | <u>0.063</u> | <u>0.003969</u> | <u>22.21</u> | <u>.365</u> | <u>0.133225</u> |
| <u>1.95</u> | <u>0.083</u> | <u>0.006889</u> | <u>22.26</u> | <u>.415</u> | <u>0.172225</u> |
| $\sum x_i = 37.34$ | $\sum (x_i - \bar{x}) = 0$ | $\sum (x_i - \bar{x})^2 = 0.183620$ | $\sum y_i = 436.9$ | $\sum (y_i - \bar{y}) = 0$ | $\sum (y_i - \bar{y})^2 = 3.2021$ |
| $\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{20}}$ $\sigma_x = 0.0958175349$ | | | $\sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{20}}$ $\sigma_y = 0.4001312285$ | | |

Figure 2: Found the sum of X and the sum of Y by square rooting the sum of x and y values and subtracting it by the mean. Then took that number and divided by the number of trials.

After performing the calculations on the calculator the numerical values turned out to be correct. The significance of performing the Standard deviation is to analyze and display how much dispersion occurred during the trials and sets of data. It shows any outliers and gives percentages of errors that exist from the average. By using the standard deviation we are able to determine the percentage of error in the linear relationship with the swimmers height and the lap times.

| $\delta xy = \overline{xy} - \bar{x} * \bar{y}$ | | | $\bar{x} = 1.867$ |
|---|-------------|---------|--|
| X Height | Y Lap times | XY | $\bar{y} = 21.845$ |
| <u>1.95</u> | 20.59 | 40.1505 | $\bar{x} * \bar{y} = 40.784615$ |
| <u>1.90</u> | 21.33 | 40.527 | $815.37/20 = 40.77685$ |
| <u>1.91</u> | 21.33 | 40.814 | $\delta xy = \overline{xy} - \bar{x} * \bar{y}$ (covariance) |
| <u>1.67</u> | 21.42 | 35.7714 | $40.77685 - 40.784615 = -.007765$ |
| <u>1.87</u> | 21.59 | 40.3733 | $\delta x^2 = 0.0958$ |
| <u>1.97</u> | 21.78 | 42.9066 | $y - 21.845 = -0.8458(x - 1.867)$ |
| <u>1.95</u> | 21.84 | 42.588 | $y - 21.845 = -0.8458(x - 1.867)$ |
| <u>1.90</u> | 21.87 | 41.553 | $y - 21.845 = -0.8458x + 1.5791$ |
| <u>1.95</u> | 21.87 | 42.6465 | $y = -0.8458x + 1.5791 + 21.845$ |
| <u>1.85</u> | 21.90 | 40.515 | $y = -0.845x + 23.4241$ |
| <u>1.90</u> | 21.91 | 41.629 | |
| <u>1.92</u> | 22.05 | 42.336 | |
| <u>1.94</u> | 22.08 | 42.835 | |
| <u>1.84</u> | 22.10 | 40.664 | |
| <u>1.80</u> | 22.15 | 39.870 | |

| | | | |
|-------------|-------|-----------------------|--|
| <u>1.70</u> | 22.14 | 37.672 | |
| <u>1.64</u> | 22.18 | 36.3752 | |
| <u>1.80</u> | 22.21 | 39.978 | |
| <u>1.93</u> | 22.21 | 42.8653 | |
| <u>1.95</u> | 22.26 | 43.407 | |
| | | $\sum xy$ =815.537 | |

Figure 3: Found covariance through the covariance formula and Pearson's product relationship formula. The formula was found from IB formula packet. The purpose for the covariance was to determine the relationship between the variables. This along with the Pearson product relationship formula aided in the process.

Based on the calculations we were able to place the standard deviation and its various calculations and I found my data to be correct. $y = -0.845x + 23.4241$. After checking my data with excel, my data turned out to be correct.

Interpretation of the results:

Based on the data calculated, there is still a correlation. Although not seen visually there seems to be a regression based on height vs. swimmer's lap time. Based on the graph itself it can be seen that the majority of swimmers with a greater height had generally faster lap times. However this statement can be made false to the outliers of some of the Japanese Swimmers such Ito Kenta who was in the top 5 swimmers with the fastest swimming times, despite his height measuring at a mere 1.67 M. Using the formula provided above, I was able to calculate the downward trend of the linear regression and what started me most was the data points shows that decreasing a person's height also decreased their free style stroke lap times. This can be stated as when looking at the linear regression line, the line passed through or exhibited points that were declining as the lap swim times went slower and slower.

However some outliers did occur that provided exceptions to the formulated hypothesis. For example, the Danish swimmer Ankjaar Jakolo was one of the people with the slowest lap times despite having a height of 1.95. This made me believe that height still does play a significant role but other factors need to be put in place such as a person's foot size, wingspan, and even diet. Even though the greater the correlation is to 1 the higher the relationship, still a relationship of 4%, concluding that height plays a trivial factor in determining the speed of a swimmer. Based on the results there was a weak positive correlation between a swimmer's height and their lap time.

Another aspect of this data was to observe the swimmer's average height, this was done through the calculation of adding all of the swimmers height and then dividing it by 20 as that was the number of swimmers, which resulted in an average height of 1.486. While the same method was used to calculate the average lap time which turned out to be 15.294, based on the graph some outliers were formed which inhibited a direct 100% correlation between height and

swim times. However a correlation of 4% indicated that the height of swimmer does matter in determining their respective lap times. Based on the data collected and graphed, it cannot be safe to generalize that swimmers who generally have a higher height will generally do better in a race. The final formula was calculated through the covariance and intercepts. The purpose for calculating the necessary covariance was to determine the relationship and points on where possible x intercepts can occur on the y intercept. As a result, by calculating the variance we found the relationships of the variance with the variables. Once the method of the strength is identified it can be determined by how great or small the covariance is to determine any linear relationship.

Standard deviation was calculated and based on the result turned out to be far greater than expected. The values yet were contradictory stating the x and y bars in the negative quartile, thereby negatively affecting the data. There is further evidence to suggest of a weak positive linear relationship and that is through the r value. Once calculated, the r value suggested that the linear relationship was so low at 0.20 that the swimmer's height and their respective lap times showed a very weak correlation between the two sets of data.

Discussion of Validity:

Errors were a major issue throughout this whole investigation. Errors began as soon as I began collecting data, initially I proposed to complete data about swimmers, heights and lap times through the 2012 Summer Olympics, yet data was incomplete and sometimes unapproachable due to inaccessibility of their premium records. However I was able to find a better alternative known as FINA which is the international governing body of swimming. By selecting their lap times I was assured these results were accurate and every swimmer in their respective countries would be able to record their data as FINA did not look at various dates, but only displayed the fastest male swim times in the world. The lap times could have been more accurate as maybe FINA could have rounded to the nearest hundredths therefore ignoring the thousandths second that could have varied the lap timings and rankings of swimmers.

Another error I encountered when performing the calculations, I was using the numbers given to me therefore there was some degree of inaccuracy especially when calculating the average height. Therefore I used the numbers I had to perform the linear correlation; however there are great limitations when judging just 20 swimmers from the myriad found around the world. Therefore more data of swimmers needed to be collected, to calculate a greater linear regression and prove to be more accurate in results. Hence, maybe in the future I would collect the world's top 50 swimmers and measure their lap times however not much information is given about swimmers who have not been given their lime light.

After performing the necessary calculations, it can be determined that linear regression was an appropriate method for investigating the solution to the problem. As based on the data collected, the weak correlation was confirmed based on the covariance formula.

Conclusion:

After completing this project, I was actually disappointed in the results of my project. There was little to none of the linear relationship between a swimmer's height and their respective lap times. Therefore my research prediction proved false. However if I had a greater yield of swimmers, I may have been able to calculate a greater linear relationship yet collecting data proved to be most cumbersome part of the whole project. When finding out the height I was citing numerous sources and used information from BBC Olympics database to find the heights. While I trusted BBC average heights could differ and thus have hindered my data collection. However despite the negative correlation I enjoyed working on this project, as I have used Math Skills to solve problems people encounter in real life but never realize. A way to improve this project would be to select greater data to plot and graph into points and possibly include other factors such as wingspan, and foot size of swimmers for the top 50 swimmers around the world. Based on that data I am sure a more accurate sense of linear relationship maybe found with additional information from encyclopedias, sport magazines, articles, and online sport databases. Evidence supported the question but the results were so minimal that no actual linear relationship could be taken into effect. I was looking for a relationship method and therefore I used linear regressions to properly validate the data.

Appendix:

Top 20 World Men's 50 meters Freestyle

| Rank | Time (seconds) | Name: (Last, First) | Date: | Team | Height (meters) |
|------|-------------------|--------------------------|----------|------|--------------------|
| 1. | 20.59 | Celio, Caesar | 20/08/12 | BRA | 1.95 |
| 2. | 21.33 | Schoeman, Roland | 16/08/12 | RSA | 1.90 |
| 3. | 21.40 | Santos, Nicholas | 20/08/12 | BRA | 1.91 |
| 4. | 21.42 | Ito, Kenta | 11/02/12 | JPN | 1.67 |
| 5. | 21.59 | Fratus, Bruno | 20/08/12 | BRA | 1.87 |
| 6. | 21.78 | Abood, Matthew | 14/07/12 | AUS | 1.97 |
| 7. | 21.84 | Magnussen, James | 19/05/12 | AUS | 1.95 |
| 8. | 21.87 | Orcechawski, Daniel | 20/08/12 | BRA | 1.90 |
| 9. | 21.87 | Cheirighnini, Marcelo | 20/08/12 | BRA | 1.95 |
| 10. | 21.90 | Less, Walter | 20/08/12 | BRA | 1.85 |
| 11. | 21.91 | Maxwell. Te Haomi | 14/07/12 | AUS | 1.90 |
| 12. | 22.05 | Dawdt, Andre | 20/08/12 | BRA | 1.92 |
| 13. | 22.08 | Mangeheira, | 20/08/12 | BRA | 1.94 |

| | | | | | |
|-----|-------|-------------------------------|----------|-----|------|
| | | Gabriel | | | |
| 14. | 22.10 | Penera, Andre | 20/08/12 | BRA | 1.84 |
| 15. | 22.15 | Marcelo, Marcos Antonio | 20/08/12 | BRA | 1.80 |
| 16. | 22.16 | Ishli, Ryo | 11/02/12 | JPN | 1.70 |
| 17. | 22.18 | Muramastu, Yoshino | 11/02/12 | JPN | 1.64 |
| 18. | 22.21 | Messias, Fellipe | 20/08/12 | BRA | 1.80 |
| 19. | 22.21 | Costa, Yuri Andrew | 20/08/12 | BRA | 1.93 |
| 20. | 22.26 | Ankjaar, Jakolo | 31/03/12 | DEN | 1.95 |

References:

1. "Fina.org - Official FINA Website." *Fina.org - Official FINA Website*. N.p., n.d. Web. 12 Sept. 2012. <http://www.fina.org/H2O/index.php?option=com_wrapper>.
2. "London 2012 Olympics - Schedule, Results, Medals, Tickets, Venues." *London2012.com*. N.p., 09 Mar. 2012. Web. 12 Sept. 2012. <<http://www.london2012.com/index-olympic.html>>.