**Internal assessment type II : Gold Medal Heights ( Answer Key)**

**(A Student work was modified to make this answer key)**

 The high jump is an Olympic sport in which a person attempts to jump over a horizontal bar without any additional aid. In this paper, I will investigate the relationship of the men’s gold medalist high jump height against the year. I will use several models and types of functions in an attempt to mirror the data as closely as possible, and also for the purpose of extrapolation of the data.

 The first data set is the men’s gold medalist high jump heights for the years 1932-1980. I set the year 1896 as 0 because 1896 was the year of the first modern Olympic Games (“Modern Olympic Games”). This means that 1932 would correspond to an x-value of 36, 1936 would correspond to an x-value of 40, and so on. Therefore, the x-axis becomes “Years after 1896”. The data is included below:

|  |
| --- |
| Men's Gold Medalist High Jump Heights |
| Years after 1896 | Height (cm) |
| 36 | 197 |
| 40 | 203 |
| 52 | 198 |
| 56 | 204 |
| 60 | 212 |
| 64 | 216 |
| 68 | 218 |
| 72 | 224 |
| 76 | 223 |
| 80 | 225 |
| 84 | 236 |

Using LoggerPro, I plotted these data points on a graph. The x-axis, Years after 1896, is the independent variable. The y-axis, Height in cm, is the dependent variable:



Legend

Plotted Points

**Parameters and Constraints:**

Since the Olympic Games were not held earlier than 1896, which has been defined to be an x-value of 0, . The Olympic Games may continue to be held for hundreds of years, so there is no upper limit to x-values.

.

It is logically impossible to have a jump height of lower than 0 cm because the competitors for the high jump are jumping upwards. Also, it is biologically improbable that humans will be able to achieve a high jump height even close to 400cm anytime within the foreseeable future. I will set 400cm as the upper limit because it is improbable that humans will ever reach this height for the high jump.

.

**Behavior of the data:**

 Below is a table of the behavior of the data:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x1 | x2 | Increase/Decrease | y1 | y2 | Slope between (x1,y1) and (x2,y2) |
| 36 | 40 | increase | 197 | 203 | 1.5 |
| 40 | 52 | decrease | 203 | 198 | -0.417 |
| 52 | 56 | increase | 198 | 204 | 1.5 |
| 56 | 60 | increase | 204 | 212 | 2 |
| 60 | 64 | increase | 212 | 216 | 1 |
| 64 | 68 | increase | 216 | 218 | 0.5 |
| 68 | 72 | increase | 218 | 224 | 1.5 |
| 72 | 76 | decrease | 224 | 223 | -0.25 |
| 76 | 80 | increase | 223 | 225 | 0.5 |
| 80 | 84 | increase | 225 | 236 | 2.75 |

 The slope is determined by 🡪 ex) 

As observed from the graph, the y-value increases, decreases for a short while, and then follows a fairly positive slope after (except from  to , where the slope is -0.25). The graph does not appear to be linear; instead, it looks like a high degree polynomial because of all of the slope changes from positive to negative and negative to positive. Because the graph appears to be behaving like a high degree polynomial, I will attempt to model the data using a quartic function.

**Model 1:**

In order to algebraically find a quartic function, I need 5 data points to solve for. I have chosen the 5 data points that seem to best represent the curvature of the graph and set up a system of 5 equations below:











Then, I set up these 5 equations as matrices in the form [A][X]=[B]. Since I am going to be solving for the coefficient matrix, [A] is the variable matrix (the x-values I am using), [X] is the coefficient matrix, and [B] is the constant matrix (the y-values I am using):



Using a TI-84 graphing calculator, I calculated the inverse of matrix *A* and then multiplied it by matrix *B*.



I used a graphing calculator to compute this operation, and got the following coefficients:



Therefore, the equation for Model 1 is:

With domain [36, 84] – from 1932 to 1980

Using LoggerPro, I graphed Model 1 with the original data points in order to see how close the model fits the data:



Legend

Model 1

Plotted Points

 The graph appears to fit the given data very well. There is a major error at, where Model 1 is much higher than the given data value of (36, 197). There are several minor errors at  and , but otherwise Model 1 fits the given data. A major limitation of Model 1 is in extrapolation of the data due to the end behavior:

As 

As 

**So Model 1 can only accurately interpolate the data to find winning heights within the years 1932 and 1980**.

 In order to test the accuracy of Model 1 to the data statistically, I will use root-mean-square deviation (referred to as RMSD).

, where # of years after 1896, = the given data’s y-value, Model 1’s y-value, and number of data points.

In this case, I was given 11 data points, so *n =* 11. I have listed the values for ,, and , and  in the table below, with values rounded to 4 significant figures for convenience (exact values were used in calculation):

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 36 | 197 | 225.6 | 817.5 |
| 40 | 203 | 203.0 | 0.0000 |
| 52 | 198 | 198.0 | 0.0000 |
| 56 | 204 | 204.5 | 0.2951 |
| 60 | 212 | 211.0 | 1.075 |
| 64 | 216 | 216.0 | 0.0000 |
| 68 | 218 | 219.2 | 1.555 |
| 72 | 224 | 221.1 | 8.133 |
| 76 | 223 | 223.0 | 0.0000 |
| 80 | 225 | 227.0 | 3.805 |
| 84 | 236 | 236.0 | 0.0000 |

Values for **** were obtained by graphing Model 1 on a graphing calculator, and then using the tabulate feature to see the (x,y) coordinate values. Since *i* = x, the corresponding y-value = ****.

Calculation for**🡪**

First I calculated all of the squares of the differences. Then I calculated the sum of the squares of the differences:

=

Now that I have the sum of the squares of the differences and *n*, I can now solve for the *RMSD*:



The lower the *RMSD* the better, but I cannot interpret it fully until I develop another model to compare it with, so I will just leave this value alone for now and come back to it later.

**Model 2:**

I will now use LoggerPro to find the quartic regression function using LoggerPro’s regression feature. Using the regression feature will most likely generate a more accurate function. The following graph shows the original plotted data, Model 1, and Model 2:



Legend

Model 1

Model 2

Plotted Pointsdeled

**Model 2: **

**With domain [36, 84] – from 1932 to 1980**

Model 1 passes through more of the data points exactly and more closely follows the data points. However, although Model 2 does not pass directly through any of the data points, it follows the general curvature of the data, and is better suited to extrapolation because the slope of the curve is not as steep. For example, at , Model 1 has a y-value of 433.4, while Model 2 has a y-value of 233.5. Both values are not very accurate because the actual winning height 24 years after 1896 was 193cm. But Model 2 is significantly closer, and therefore is better for extrapolation of the data. A comparison of the *RMSD* values shows that Model 2 is indeed the more accurate model:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 36 | 197 | 200.0 | 9.137 |
| 40 | 203 | 197.4 | 30.88 |
| 52 | 198 | 202.4 | 19.30 |
| 56 | 204 | 206.1 | 4.259 |
| 60 | 212 | 210.0 | 4.089 |
| 64 | 216 | 213.9 | 4.508 |
| 68 | 218 | 217.6 | 0.1317 |
| 72 | 224 | 221.3 | 7.453 |
| 76 | 223 | 224.9 | 3.690 |
| 80 | 225 | 228.9 | 14.98 |
| 84 | 236 | 233.5 | 6.076 |



Model 1 had a *RMSD* of 8.699, so Model 2 has a much lower *RMSD* and would therefore be a more accurate model of the data from a statistical viewpoint. However, looking at the *RMSD* calculations for Model 1, the square of the differences atwas 817.5. This large number greatly skewed the sum of the squares of the differences, which was only 832.4. Therefore, *RMSD* cannot be the sole factor for determining which function models the data more accurately. As stated before, Model 1 interpolates the data better than Model 2 and Model 2 extrapolates the data better than Model 1. Every graph has its strengths and weaknesses, and Model 2’s strength is Model 1’s weakness (and vice versa).

Therefore, Model 1 will be used for interpolation of the data, and Model 2 will be used for extrapolation of the data. **(Notes: Both models would not give a good results for extrapolation predicting the winning heights of 1984 and especially 2016)**

**Winning Heights in 1940 and 1944:**

 These years are within the range of the data set, so I will interpolate and use Model 1 to determine what the winning heights would have been. Inputting the function for Model 1 and tabulating the results, I get the following data:

|  |  |  |
| --- | --- | --- |
| Years after 1896 | 44 | 48 |
| Height (cm) | 193.8 | 193.4 |

The year 1940 corresponds to , which has a winning height of 193.8cm. When, or the year 1944, the winning height is 193.4cm. These data points seem to be logically correct because World War II occurred during this time period, so athletes would probably not be training as much as they had been in previous years. Therefore, the winning height would have dropped from 1936 to 1940 because of lack of training, and then another drop from 1940-1944 would be expected, again because of the lack of preparation for the Olympics.

**Winning Heights in 1984 and 2016:**

 I will use Model 2 to predict the winning heights in 1984 and 2016 because this involves extrapolation of the data. Using a graphing calculator, I graphed Model 2 and tabulated the results for the x and y-values corresponding to the years 1984 and 2016:

|  |  |  |
| --- | --- | --- |
|  **Years after 1896** | 88 | 120 |
| **Height (cm)** | 239.5 | 442.9 |

 The predicted winning height for 1984 is an acceptable value because the generally positive trend in the data suggests that improvement would be likely. However, the predicted winning height for 2016 is 442.9cm, which is not within the parameter for y-values. I set the maximum value for the winning height at 400cm because it is biologically improbable for humans to jump even close to 400cm. Therefore, this predicted value of 442.9cm is not acceptable. As a result, Model 2 has been shown to be useful for extrapolating only a few years past the data set. After that, the predicted values become too high to be logically acceptable.

**Including Data for All Olympic Games since 1896:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Years after 1896** | 0 | 8 | 12 | 16 | 24 | 32 | 88 | 92 | 96 | 100 | 104 | 108 | 112 |
| **Height (cm)** | 190 | 180 | 191 | 193 | 193 | 194 | 235 | 238 | 234 | 239 | 235 | 236 | 236 |

Using LoggerPro, I have graphed the additional data points with the original data. Because these data points are beyond the original data set, I will use Model 2 to see how well it fits the additional data:



Legend

Model 2

Plotted Pointsdeled

Model 2 (**)**  does not fit the additional data at all. The additional data points are much lower than the values predicted by the model. Around 90 years after 1896, Model 2 begins to deviate from the data, going higher and higher, while the actual data starts to level off at the same time.

 The overall trend from 1896 to 2008 is a steady increasing winning height. There are several fluctuations throughout the data set, such as from 1896 to 1908 and from 1936 to 1948, where the height for the high jump goes down and then starts to go back up.

 In order to fit the new data, a different type of function must be used. The end behavior of this function must have a defined upper and lower limit in order to fulfill the parameter for the range of y-values:



 Logistic functions have both an upper and lower limit, and so I will use a logistic function to model the data.

The domain of x was originally set from zero to positive infinity because the modern Olympics were not held earlier than 0 years after 1896. This will not affect my choice of function – the function will still be able to calculate values earlier than 0 years after 1896; however, these values will be thrown out because there were no modern Olympic Games before 1896.

Logistic functions are written in the form , where is the value of the horizontal asymptote for the upper limit, *e* is Euler’s number, 2.71828183, *d* is the value of the horizontal asymptote for the lower limit, and *b* and *c* are coefficients which modify the graph. I have set the lower limit for y-values at 0cm, so . I have set the upper limit for y-values at 400cm, and since the lower limit is 0cm, I can substitute 400cm for *a*.



Using LoggerPro’s regression feature, I created the following logistic regression function:



**Suggest to formulate this equation algebraically as well**.

 I have graphed the logistic function, along with Model 2 and all of the data points in order to determine the better model:



Legend

Model 2

Model 3

Plotted Pointsdeled

Model 3 represents the general trend in the data and fits the data better than Model 2. There is no significant deviation from the data for Model 3, while Model 2 experiences significant deviation on the far left and far right of the graph. Zooming out, the end behavior of these two models can be compared:

**RMSD value needs to be calculated for Model 3 function (should be about 4). Discuss between the Model 2 function and Model 3 function comparing RMSD values and their end behaviors of the functions.**

**Also using Model 3, recalculate the predicated heights of 1984 and 2016.**



Legend

Model 2

Model 3

Plotted Pointsdeled

Model 3’s end behavior fits the parameter stated earlier: 

As  and as 

Model 2 does not fulfill the parameter for y-values; instead, as , . Model 3 represents the data better and can extrapolate the data more accurately due to the end behavior. Therefore, Model 3 is the best model of the data and will be used for the rest of this investigation.

**Comparing Model 3 to other sets of data:**

Below is data for the women’s high jump gold medalist heights since 1896:

|  |
| --- |
| Women's Gold Medalist High Jump Heights |
| Years after 1896 | Height (cm) |
| 32 | 159 |
| 36 | 165 |
| 40 | 160 |
| 52 | 168 |
| 56 | 167 |
| 60 | 176 |
| 64 | 185 |
| 68 | 190 |
| 72 | 182 |
| 76 | 192 |
| 80 | 193 |
| 84 | 197 |
| 88 | 202 |
| 92 | 203 |
| 96 | 202 |
| 100 | 205 |
| 104 | 201 |
| 108 | 206 |
| 112 | 205 |

(Rosenbaum, “High Jump Women's Olympic Medalists”)

Using LoggerPro, I have graphed the data for the women’s high jump with Model 3 to test Model 3’s applicability to other sets of data:

**Use regression feature and find a new function (logistic) to model these data. The best fit logistic function is: **



Legend

Model 3

Plotted Pointsdeled

The data for the women’s high jump is significantly lower than Model 3 – none of the points even come close to the model. The women’s data looks like it may not even fit into a logistic function; instead, a linear function would also work for this data set. There are still some fluctuations in the women’s data, but the general trend is a steady, positive slope.

If a logistic function was used to model this data set, then the variable would have to be lower in order to have a smaller value for the upper limit. The women’s data could fit into a logistic function by translating Model 3  units down. After this, the end behaviors would have to be adjusted (because a downward translation of 40 units would put the lower limit at -40 cm, which is outside of the parameter for y-values.

**Discuss the result with the best fit logistic function found.**

**Conclusion:**

After finding three functions to model the data, Model 3 was the most accurate. It represented the general trend in the data fairly well, but most importantly, it fulfilled the parameter for the y-values, which the other models were unable to do. Also, it does not reach the upper limit too quickly. For example, 500 years after 1896, the year 2396, Model 3 predicts a winning height of 371.3cm. This number is still very high and unlikely according to our current knowledge of our biological capabilities, but things may be much different 400 years from now. Therefore, a winning height of 371.3cm is still an acceptable value.

* **A new function,  may prove somewhat accurate values in predicting the winning height for the women’s high jump.**

In addition, the data itself had some irregularity in the winning height for the men’s high jump. As shown from the men’s high jump gold medalist heights, the winning height often fluctuated (from 190cm in 1896 to 180cm in 1904 and then back to 191cm in 1908). This can be explained by the frequency of the summer Olympic Games – they are only held every 4 years. This is a very low frequency, so one bad day for the top competitor may result in a lower winning height. Competitors must also be consistent in the high jump – three consecutive failed attempts and they will be eliminated from the competition (Rosenbaum, “What is Olympic High Jump?”). Also, an increase in the winning height is not necessarily due to an advance in athletic performance. The high jump involves complex techniques that must be executed perfectly in order to clear the bar (Rosenbaum, “Illustrated High Jump Technique”). Therefore, using better techniques can lead to improvements in the winning height for the high jump. All of these factors can affect the data set and therefore affect the model functions that would be used to extrapolate the data.

But for the most part Model 3, , does an excellent job of modeling the given data (from 1896 to 2008) , and it is a reasonable model for this investigation. Therefore, a logistic function can be used to model the winning heights of the men’s high jump, and is a fairly accurate modeling of the data.

Works Cited

 “Modern Olympic Games.” *Infoplease*. Pearson Education, 01 Feb 2009. Web. 4 Dec 2011.

Rosenbaum, Mike. “Illustrated High Jump Technique.” *About.com*. About.com, 2011. Web. 4 Dec 2011.

Rosenbaum, Mike. “High Jump Women's Olympic Medalists.” *About.com*. About.com, 2011. Web. 10 Dec 2011.

Rosenbaum, Mike. “What is Olympic High Jump?” *About.com*. About.com, 2011. Web. 11 Dec 2011.