Chunyang Ding

Mr. Kessler

AP/IB Calculus Mathematics SL

14 December 2012

Gold Medal Modeling Portfolio

The year is 2016, and you are an American long jumper contesting to compete in the Rio de Janerio Olympics. You know that regardless of what happens, you want to do your very bets, but it sure would be really comforting if you were able to predict the height that would net you the gold medal. When it comes down to the moment, when the entire country is watching you, and when the world focuses their eyes on your final Fosbury Flop, you need to be one hundred percent prepared, both physically and mentally. You HAVE to live in that moment; it is the only way you will walk out with success.

One of the ways that any high jumper can better prepare themselves is by knowing the competition, or at least knowing what the competition should be able to do. Unfortunately, all previous attempts to spy on the Russian and Chinese team's practices have led to a speedy eviction, usually with several large dogs following behind. Another way that you could potentially outmatch or outwit the competition is by studying the trends of Olympic high jumps, and from that, figure out the minimum height that you should be reaching.

To do so, you find data of previous Olympic high jump records from the International Olympic Committee, and begin processing the data. In order to predict the gold-medal height for this year, one should try to find a general trend of the data, and use that model. From the model should emerge the general trend of the high jumps through the year, revealing a better understanding of the evolution of the event.

Initially, the only data you find is from your coach's old notebooks, dating from 1932 to 1980. This data is shown below, with a slight modification. In order to simplify the regression process for the trend, it is better to not take the entire year in account, but only the number of years passed since 1900. This provides a good baseline for your model and would be easy to understand.

Years since 1900	Height of gold medal in cm
32	197
36	203
48	198
52	204
56	212
60	216
64	218
68	224
72	223
76	225
80	236

Figure 1.1

The first thing that we notice is that as the years progressed, the gold-medal height also increased. This shows a clear positive correlation between the years passed and the height achieved. Thinking about the event, you realize that this does in fact make sense. Every year,

competitors come out of the event with new ideas of how to train and better ways to safely improve their body. Everybody wants to beat the last year's record, and would train as much as possible to do so.

In order to better visualize what kind of trend existed between the data, we plot the data with the years on the x-axis and the maximum height on the y-axis, as we believe that the number of years does influence the maximum height.

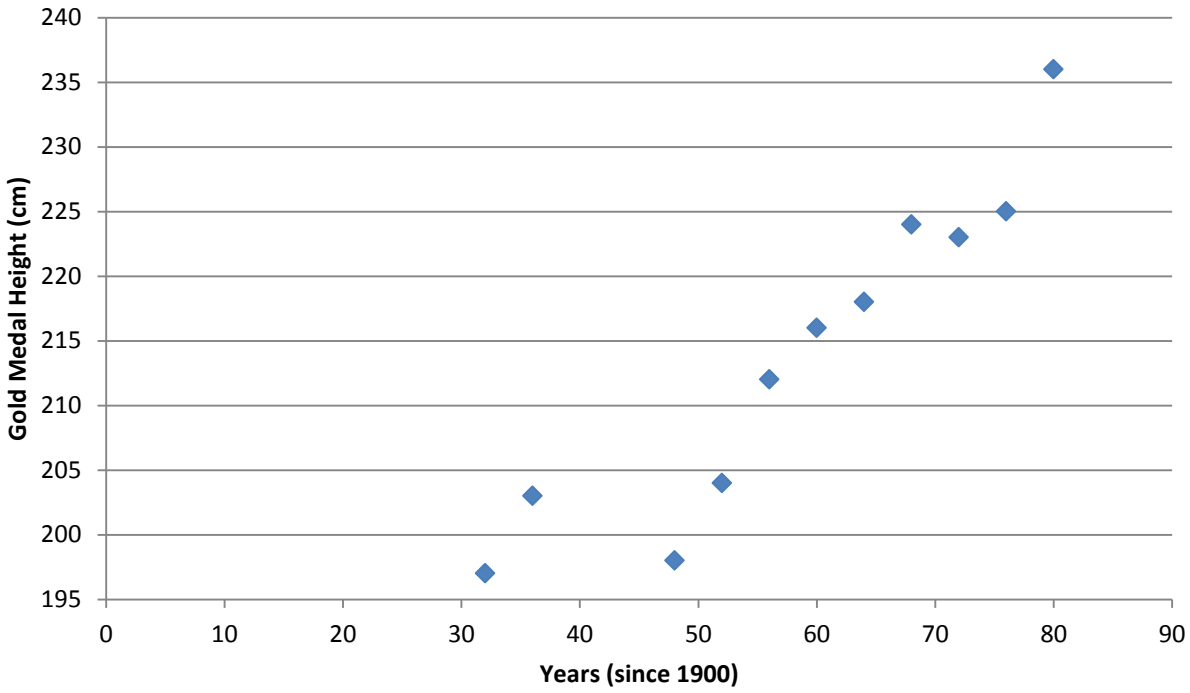


figure 1.2

In order to determine what kind of line would best fit the data provided, all sorts of functions should be looked at. There are many families of functions, but because the positive correlation has already been identified, we can eliminate several families, such as any inverse functions or sinusoidal functions. The most likely functions remain as either linear, quadratic, or

power functions. Each case should be studied and compared in order to reason which would most fit the data, as well as which one would make the most sense.

The easiest function to model would be a linear function, in the form

$$f(x) = mx + b$$

where m represents the slope of the line and b is the y-intercept. By just looking at the slope between the first and last points, we can easily find a rough estimate for the slope of the entire graph. As slope is found by

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

and our points are $(32, 197) and\ (80, 236)$ for the minimum and maximum heights, respectively, we can reason that the slope of the best fit line should be

$$m = \frac{236 - 197}{80 - 32}$$

$$m = \frac{39}{48}$$

$$m = 0.8125$$

While this is by no means a perfect slope, it does provide us with a good estimate for the actual slope of the best fit line.

Using the information, our equation has only one more variable left: the y-intercept. Finding this variable is extremely easy, as the only step required is to substitute any data point into the partial equation and solve for the b variable. Using the point $(60, 216)$, we can discover:

$$f(x) = 0.8125x + b$$

$$216 = 0.8125\,(60) + b$$

$$216 - 48.75 = b$$

$$167.25 = b$$

Now, we have a full equation for the approximate line of best fit, as shown below alongside with the data

$$f(x) = 0.8125x + 167.25$$

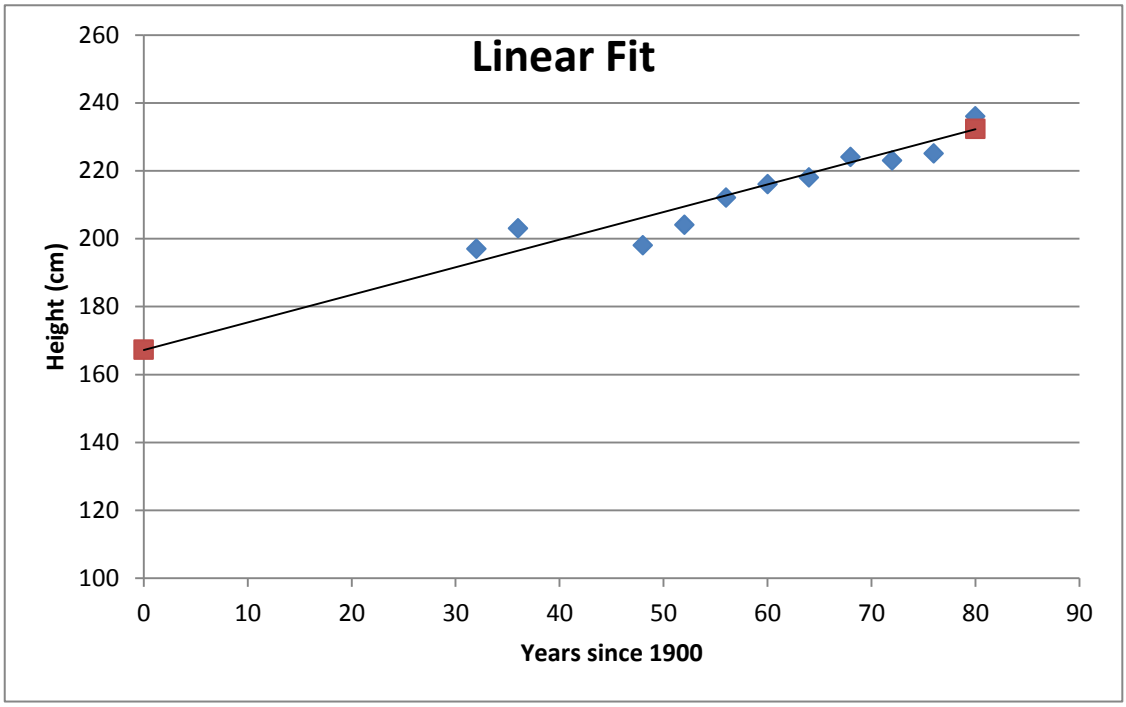
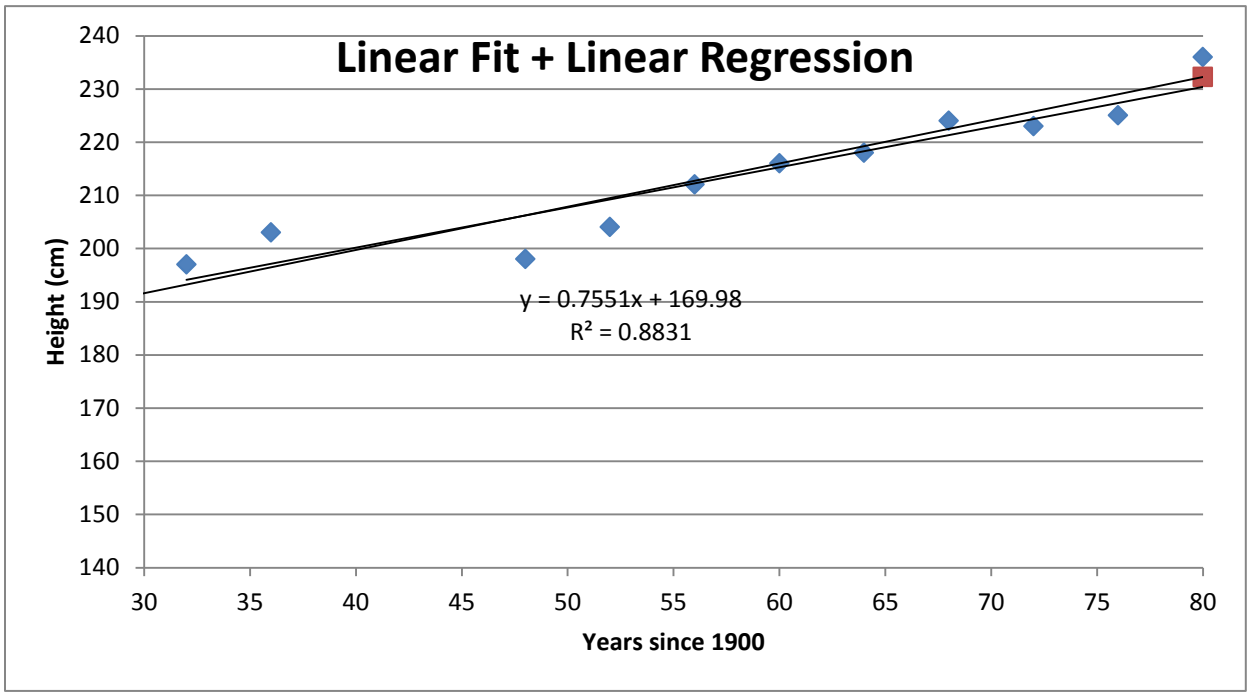


figure 1.3

Our graph is not extremely accurate, as there are many points both above and below the line of best fit. However, we can see that this line is relatively close and models the correct

correlation of data. Using our technology, we are able to generate the actual line of best fit, and we can compare the two lines side by side, as seen below.



Insert figure 1.4 *here*

As you can see from the graph, there is extremely little difference between the two trend lines, even though the equations of the actual regression, $f(x) = 0.7551x + 169.98$, are off by a bit. This goes to show that our original analytical regression is not all that much off of the correct regression.

However, we could tell from the graph that the line is likely not the best fit possible. There seems to be a considerable area around the start of the graph that is below our best fit line, as well as several points that seem to hover above the line of best fit. Later on, we will evaluate

exactly how well does our data fit, but for now, let us explore a different type of function that may better model our function.

It seems that it is possible that our graph actually displays a trend in the form of $f(x) = ax^n$. However, without a firm foundation in statistics, it would be incredibly difficult for us to regress that by hand properly. Currently, at this level, we do not have the necessary tools in order to work out this regression. However, programs such as Excel or Mathematica do have that capacity, and when the data is imputed, the following power graph emerges.

$$f(x) = 103.87 \cdot x^{0.1797}$$

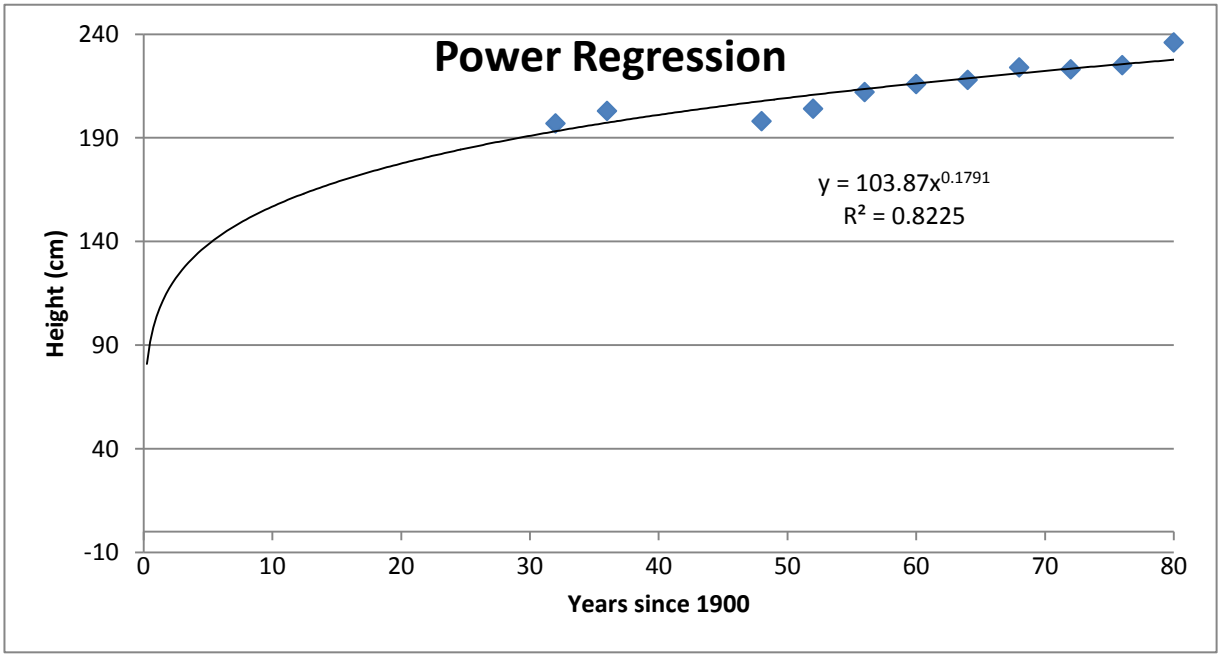


figure 1.5

One additional model that we could compare is a quadratic model. Even though this seems to be somewhat similar to the power regression, it may in fact be able to better model the data, given that there is a slight upward curve in the current data. This matches with the parabolic

shape of a quadratic graph, so we can try to apply this regression. Excel reveals the following function and graph.

$$f(x) = 0.0113 \cdot x^2 - 0.5057x + 202.67$$

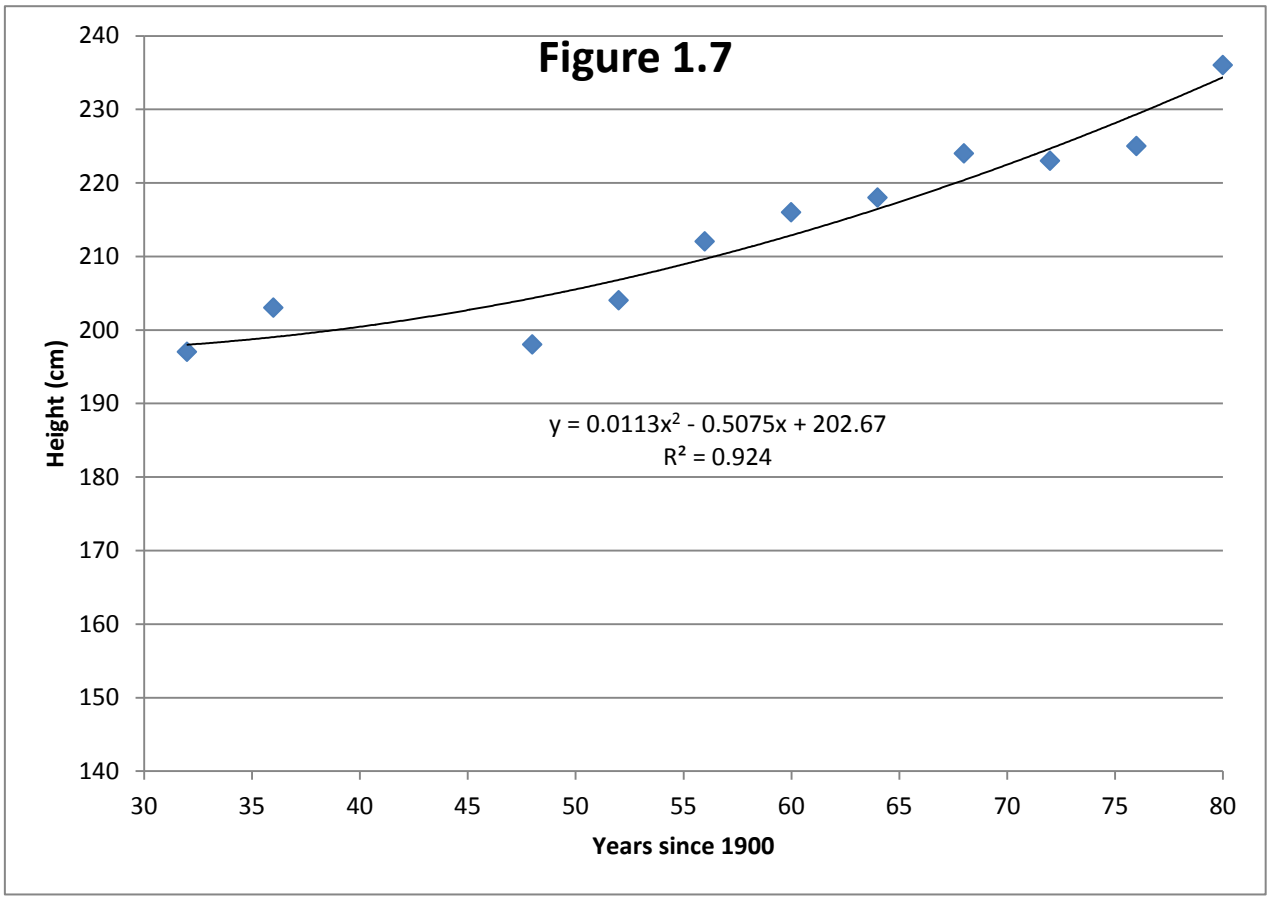


figure 1.6

Finally, another model that we could potentially test is a logarithmic model, which tends to taper off as the x values grow larger. This model could potentially make sense in the real world, because it doesn't make very much logical sense for human jumping patterns to continually grow; eventually we will reach a limit to what our bodies are capable of. Through use of Excel, the following function and graph are produced.

$$f(x) = 38.194 \cdot \ln(x) + 60.158$$

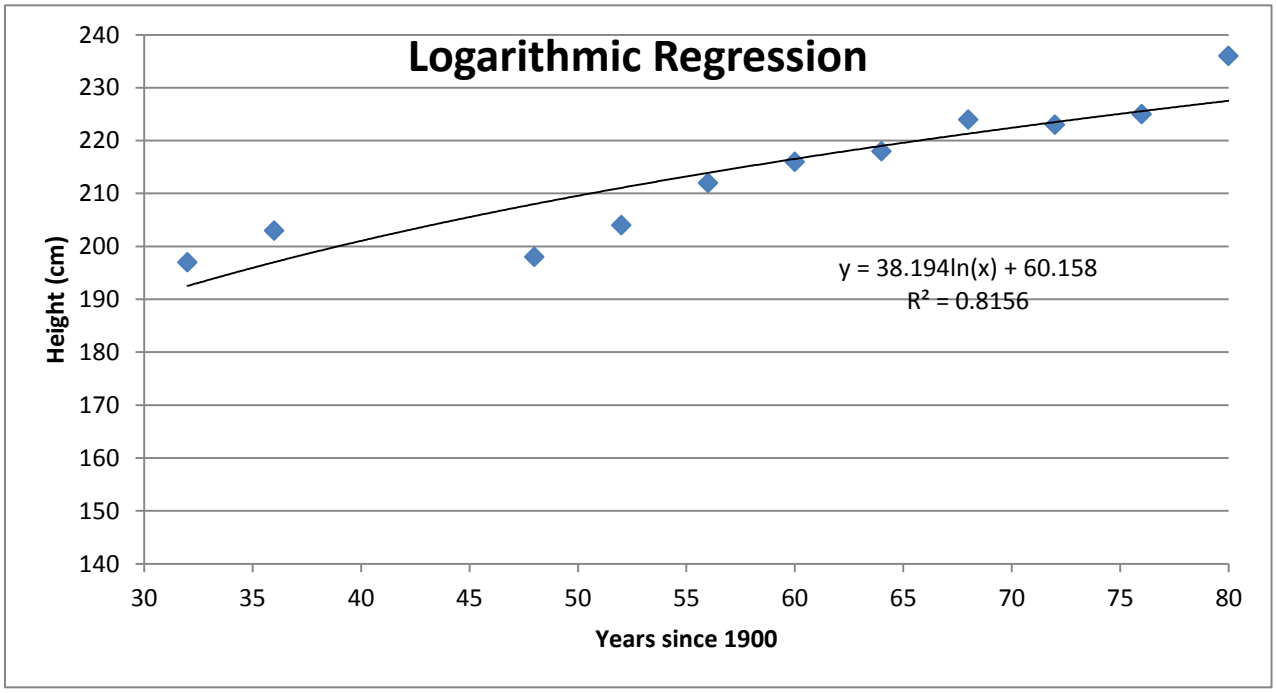


figure 1.6

Currently, we can overlay all four of the previous graphs in order to interpret which regression may best match the function. However, when we do so in figure 1.7, we may notice that all of the functions seem to match the region quite well. They generally have about the same number of data points above the curve as below the curve. Therefore, in statistics, the best way to calculate how well our function matches our data is to find the R squared value. This process isn't especially easy, but it is possible through the following steps:

$$R^2 = 1 - \frac{SSErr}{SSTot}$$

$$SSErr = \sum_{a}^{i} \left(y_a - f(i)\right)^2$$

$$SSTot = \sum_{a}^{i} (y_a - \bar{y})^2$$

$$\bar{y} = \frac{1}{n} \sum_{a}^{n} y_a$$

Even though this process is quite convoluted, it is possible to process our data from these standards. However, for the sake of accuracy, it will be better to process the R squared term with Excel. This reveals that the quadratic regression is the best way to fit the data, with an R squared value of 0.924.

However, the information provided by the quadratic equation does not truly make sense. It would imply that as time progresses, the height that people can actually jump would increase, and that increase in height would also grow. Eventually, people will be able to jump over 3 story buildings, or otherwise jump to ridiculous heights. Therefore, for practical purposes, the quadratic fit and the linear fit, to some extent, does not always make sense. However, for the small time period that we are observing, it is possible to use these regressions to predict the heights for years in the same neighborhood as the ones we currently look at.

Using the most accurate model, the quadratic model, it is possible to predict with some accuracy what the gold medal heights for 1940 and 1944 were. Our x-variable would be 40 and 44 respectively, and would yield the results:

$$f(40) = 0.0113 \cdot (40)^2 - 0.5057(40) + 202.67$$

$$f(x) = 200.522$$

$$f(44) = 0.0113 \cdot (44)^2 - 0.5057(44) + 202.67$$

$$f(44) = 202.296$$

These are reasonable results, as they are increasing, as much of the rest of the data would indicate for them to be, and they do not seem to be increasing extremely rapidly, skewing the rest of the graph.

Later on, we find more information about other gold medal trials in the Olympics. These new data points allow us to refine our original model, as if we have more data, our model should be able to match it. Shown below is previous quadratic model overlaid on the new data, as well as the new trend line that Excel has processed for us.

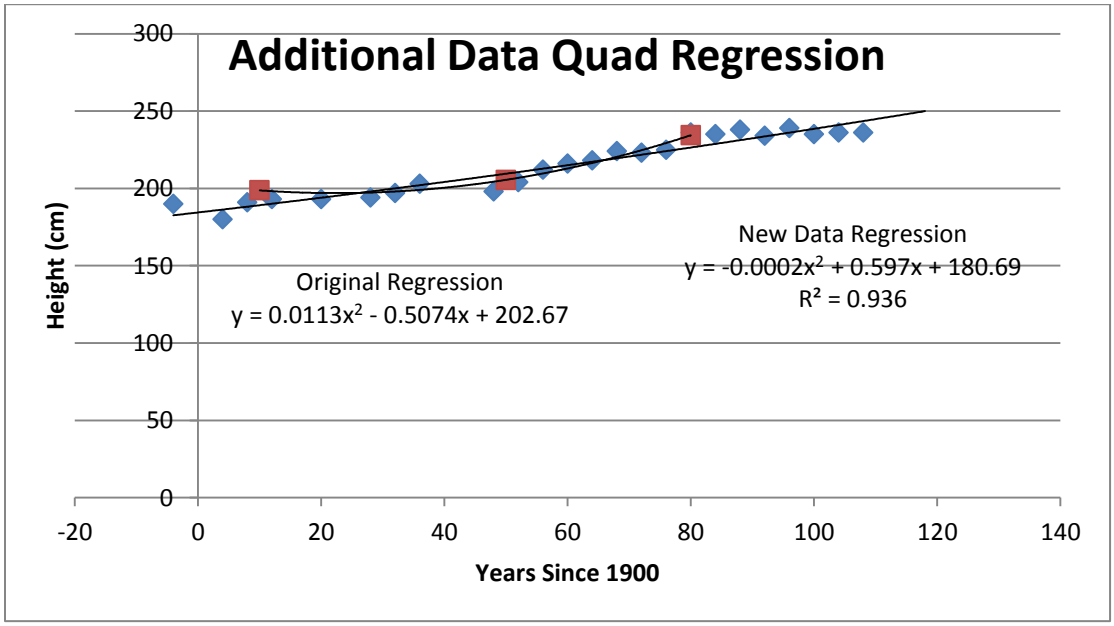


figure 1.7

As you can clearly tell, there is an extremely large discrepancy with what our previous best fit line is as compared to the current best fit line. For starters, our best fit line has a parabola

opening up, while this new parabola is opening downwards. It may be especially confusing as to why we had such a large difference, but when we take into account the minimal amount of data we began with, as well as the difficulty inherent with modeling human behaviors, the best fit lines seem to make more sense.

The new line of best fit also makes more sense to us, as it reveals the slowing of growth as years go on. Although it seems to reveal that eventually, there will be a maximum point after which humans aren't able to jump any higher, and in fact start to lose jumping ability, it may in fact point towards the gradual leveling off of the heights, which makes sense in a physical way.

Using this new graph, we again predict the heights for the "missing years" of 1940 and 1944, as well as the year 2016, as that was the original intent of this portfolio. Doing, so, we get the following calculations:

$$f(40) = -0.0002(40)^2 + 0.597(40) + 180.69$$

$$f(40) = 204.25$$

$$f(44) = -0.0002(44)^2 + 0.597(44) + 180.69$$

$$f(44) = 206.57$$

$$f(116) = -0.0002(116)^2 + 0.597(116) + 180.69$$

$$f(116) = 247.25$$

As we can tell, there are large discrepancies between our original estimates and our new estimates, but as they are still within a remarkably close range, it is safe to conclude that our new model does work to some extent.

Through looking at the data, the idea of modeling real-world data has been explored. However, it must be kept in mind that these functions do not make perfect sense. In the real world, especially in events with as much error as a high jump, it is extremely difficult to find a mathematical justification for patterns. For example, if a new type of jumping was invented, or if better shoes were created, the data would immediately be skewed based on those variables. Also, depending on a certain person's body composition or genetic discrepancies within people, there may be sudden increases due to genetic benefits, or other skews within the data. Additionally, the Olympic results are more of a result of how much people train than to how the data has shown. A possible example within our data is the odd discrepancy that occurs at 1948. This may have occurred because the athletes have not competed for a large number of years, or that new athletes did not have the needed experience at the Olympics to do well. Whatever the reason, it creates a problem for the regression, and can largely skew the data.

In reality, there are many variables limiting the height people could possibly jump to. These variables would not make sense to just increase as time progresses, as that would again imply that eventually people will be able to jump and fly. There should be an eventual leveling off, as seen in our logarithmic graphs. Therefore, any graph would eventually fail in its predictive power. However, for a region surrounding the region being regressed, we can be reasonably certain that this model would hold, thereby giving the user a reasonable guess for the gold medal heights.

With all of this mathematical work done, you have found the expected height of the gold medal height of the 2016 Summer Olympics to be 247.25 cm. Determined, you set yourself on a strict training regiment, and when the time comes, succeed in doing so well that you actually go over the expected value, and hit 253.16 cm. Although you might shrug and think about how no model is 100% accurate, as it can be so easily influenced by a variety of human factors, this is not the time for that. Instead, it is the time to revel in your success of taking home the gold medal!