Curve-fitting Project - Linear Model (due at the end of Week 5)

Instructions

For this assignment, collect data exhibiting a relatively linear trend, find the line of best fit, plot the data and the line, interpret the slope, and use the linear equation to make a prediction. Also, find r^2 (coefficient of determination) and r (correlation coefficient). Discuss your findings. Your topic may be that is related to sports, your work, a hobby, or something you find interesting. If you choose, you may use the suggestions described below.

A **Linear Model Example** and Technology Tips are provided in separate documents.

Tasks for Linear Regression Model (LR)

(LR-1) Describe your topic, provide your data, and cite your source. Collect at least 8 data points. **Label appropriately. (Highly recommended: Post this information in the Linear Model Project discussion as well as in your completed project. Include a brief informative description in the title of your posting.** Each student must use different data.)

The idea with the discussion posting is two-fold: (1) To share your interesting project idea with your classmates, and (2) To give me a chance to give you a brief thumbs-up or thumbs-down about your proposed topic and data. Sometimes students get off on the wrong foot or misunderstand the intent of the project, and your posting provides an opportunity for some feedback. Remark: **Students may choose similar topics, but must have different data sets.** For example, several students may be interested in a particular Olympic sport, and that is fine, but they must collect different data, perhaps from different events or different gender.

(LR-2) Plot the points (x, y) to obtain a scatterplot. Use an appropriate scale on the horizontal and vertical axes and be sure to label carefully. Visually judge whether the data points exhibit a relatively linear trend. (If so, proceed. If not, try a different topic or data set.)

(LR-3) Find the line of best fit (regression line) and graph it on the scatterplot. State the equation of the line.

(LR-4) State the slope of the line of best fit. Carefully interpret the meaning of the slope in a sentence or two.

(LR-5) Find and state the value of r^2, the coefficient of determination, and r, the correlation coefficient. Discuss your findings in a few sentences. Is r positive or negative? Why? Is a line a good curve to fit to this data? Why or why not? Is the linear relationship very strong, moderately strong, weak, or nonexistent?

(LR-6) Choose a value of interest and use the line of best fit to make an estimate or prediction. Show calculation work.

(LR-7) Write a brief narrative of a paragraph or two. Summarize your findings and be sure to mention any aspect of the linear model project (topic, data, scatterplot, line, r, or estimate, etc.) that you found particularly important or interesting.

You may submit all of your project in one document or a combination of documents, which may consist of word processing documents or spreadsheets or scanned handwritten work, provided it is clearly labeled where each task can be found. Be sure to include your name. Projects are

graded on the basis of completeness, correctness, ease in locating all of the checklist items, and strength of the narrative portions.

Here are some possible topics:

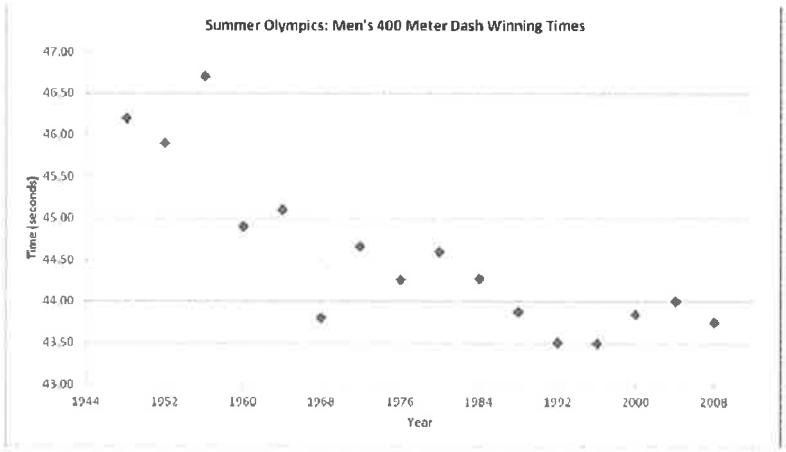
- Choose an **Olympic sport** – an event that interests you. Go to http://www.databaseolympics.com/ and collect data for winners in the event for at least 8 Olympic games (dating back to at least 1980). (Example: Winning times in Men's 400 m dash). Make a quick plot for yourself to "eyeball" whether the data points exhibit a relatively linear trend. (If so, proceed. If not, try a different event.) After you find the line of best fit, use your line to make a prediction for the next Olympics (2014 for a winter event, 2016 for a summer event).
- Choose a particular type of **food**. (Examples: Fish sandwich at fast-food chains, cheese pizza, breakfast cereal) For at least 8 brands, look up the fat content and the associated calorie total per serving. Make a quick plot for yourself to "eyeball" whether the data exhibit a relatively linear trend. (If so, proceed. If not, try a different type of food.) After you find the line of best fit, use your line to make a prediction corresponding to a fat amount not occurring in your data set.) Alternative: Look up carbohydrate content and associated calorie total per serving.
- Choose a **sport** that particularly interests you and find two variables that may exhibit a linear relationship. For instance, for each team for a particular season in baseball, find the total runs scored and the number of wins. Excellent websites: http://www.databasesports.com/ and http://www.baseball-reference.com/

Scatterplots, Linear Regression, and Correlation

When we have a set of data, often we would like to develop a model that fits the data.

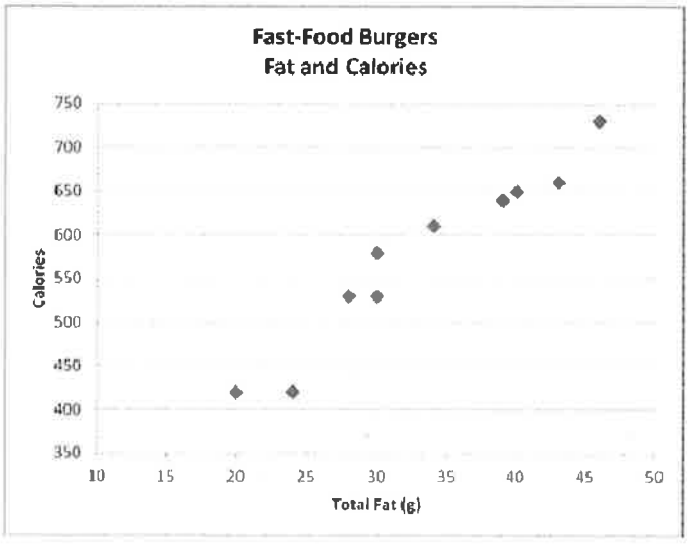
First we graph the data points (x, y) to get a **scatterplot.** Take the data, determine an appropriate scale on the horizontal axis and the vertical axis, and plot the points, carefully labeling the scale and axes.

Summer Olympics: Men's 400 Meter Dash Winning Times	
Year (x)	Time(y) (seconds)
1948	46.20
1952	45.90
1956	46.70
1960	44.90
1964	45.10
1968	43.80
1972	44.66
1976	44.26
1980	44.60
1984	44.27
1988	43.87
1992	43.50
1996	43.49
2000	43.84
2004	44.00
2008	43.75



Burger	Fat (x) (grams)	Calories (y)
Wendy's Single	20	420
BK Whopper Jr.	24	420
McDonald's Big Mac	28	530
Wendy's Big Bacon Classic	30	580
Hardee's The Works	30	530
McDonald's Arch Deluxe	34	610
BK King Double Cheeseburger	39	640
Jack in the Box Jumbo Jack	40	650
BK Big King	43	660
BK King Whopper	46	730

Data from 1997



If the scatterplot shows a relatively linear trend, we try to fit a linear model, to find a line of best fit.

We could pick two arbitrary data points and find the line through them, but that would not necessarily provide a good linear model representative of all the data points.

A mathematical procedure that finds a line of "best fit" is called **linear regression**. This procedure is also called the method of least squares, as it minimizes the sum of the squares of the deviations of the points from the line. In MATH 107, we use software to find the regression line. (We can use Microsoft Excel, or Open Office, or a hand-held calculator or an online calculator --- more on this in the Technology Tips topic.)

Linear regression software also typically reports parameters denoted by r or r^2.

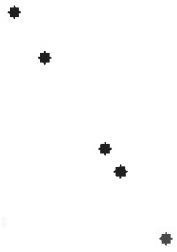
The real number ***r*** is called the **correlation coefficient** and provides a measure of the strength of the linear relationship.

r is a real number between –1 and 1.

$r = 1$ indicates perfect positive correlation --- the regression line has positive slope and all of the data points are on the line.

$r = -1$ indicates perfect negative correlation --- the regression line has negative slope and all of the data points are on the line

Correlation $r = 1$

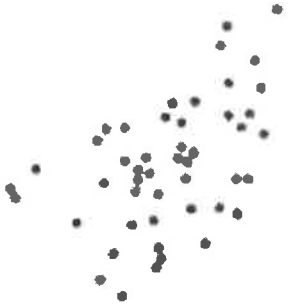


Correlation $r = -1$

The closer $|r|$ is to 1, the stronger the linear correlation. If $r = 0$, there is no correlation at all. The following examples provide a sense of what an r value indicates.

Correlation $r = 0$

Correlation $r = -0.3$



Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

Source: *The Basic Practice of Statistics*, David S. Moore, page 108.

Notice that a positive r value is associated with an increasing trend and a negative r value is associated with a decreasing trend. The strongest linear models have r values close to 1 or close to -1.

The nonnegative real number r^2 is called the **coefficient of determination** and is the square of the correlation coefficient r.

Since $0 \le |r| \le 1$, multiplying through by $|r|$, we have $0 \le |r|^2 \le |r|$ and we know that $-1 \le r \le 1$.
So, $0 \le r^2 \le 1$. The closer r^2 is to 1, the stronger the indication of a linear relationship.
Some software packages (such as Excel) report r^2, and so to get r, take the square root of r^2 and determine the sign of r by observing the trend (+ for increasing, – for decreasing).

(Sample) Curve-Fitting Project - Linear Model: Men's 400 Meter Dash

Submitted by Suzanne Sands

(LR-1) **Purpose**: To analyze the winning times for the Olympic Men's 400 Meter Dash using a linear model

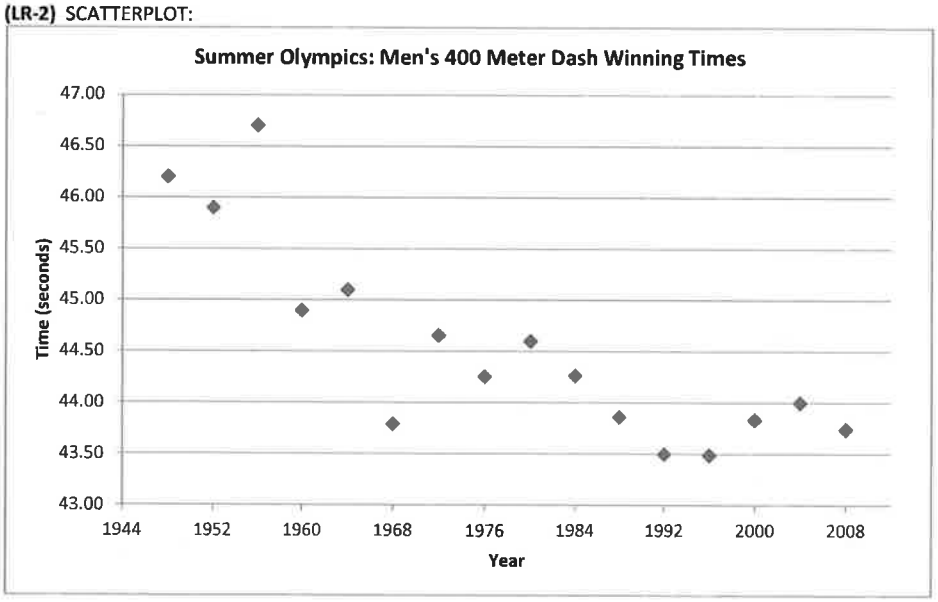
Data: The winning times were retrieved from http://www.databaseolympics.com/sport/sportevent.htm?sp=ATH&enum=130

The winning times were gathered for the most recent 16 Summer Olympics, post-WWII. (More data was available, back to 1896.)

DATA:

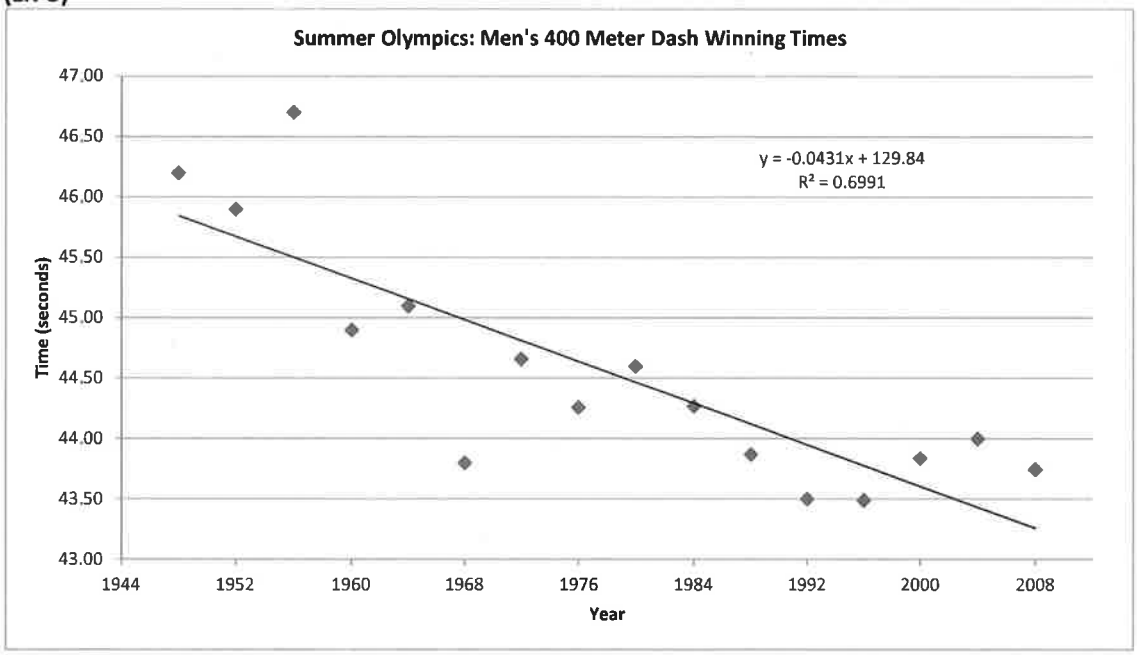
Summer Olympics: Men's 400 Meter Dash Winning Times	
Year	Time (seconds)
1948	46.20
1952	45.90
1956	46.70
1960	44.90
1964	45.10
1968	43.80
1972	44.66
1976	44.26
1980	44.60
1984	44.27
1988	43.87
1992	43.50
1996	43.49
2000	43.84
2004	44.00
2008	43.75

(LR-2) SCATTERPLOT:



As one would expect, the winning times generally show a downward trend, as stronger competition and training methods result in faster speeds. The trend is somewhat linear.

(LR-3)



Line of Best Fit (Regression Line)

$y = -0.0431x + 129.84$ where x = Year and y = Winning Time (in seconds)

(LR-4) The slope is –0.0431 and is negative since the winning times are generally decreasing.
The slope indicates that in general, the winning time decreases by 0.0431 second a year, and so the winning time decreases at an average rate of 4(0.0431) = 0.1724 second each 4-year Olympic interval.

(LR-5) Values of r^2 and r:

$r^2 = 0.6991$

We know that the slope of the regression line is negative so the correlation coefficient r must be negative.

$$r = -\sqrt{0.6991} = -0.84$$

Recall that $r = -1$ corresponds to perfect negative correlation, and so $r = -0.84$ indicates moderately strong negative correlation (relatively close to -1 but not very strong).

(LR-6) Prediction: For the 2012 Summer Olympics, substitute $x = 2012$ to get $y = -0.0431(2012) + 129.84 \approx 43.1$ seconds.

The regression line predicts a winning time of 43.1 seconds for the Men's 400 Meter Dash in the 2012 Summer Olympics in London.

(LR-7) Narrative:

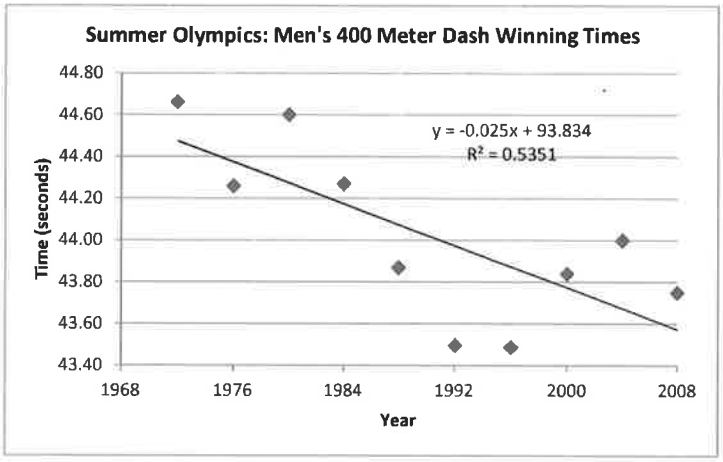
The data consisted of the winning times for the men's 400m event in the Summer Olympics, for 1948 through 2008. The data exhibit a moderately strong downward linear trend, looking overall at the 60 year period.

The regression line predicts a winning time of 43.1 seconds for the 2012 Summer Olympics, which would be nearly 0.4 second less than the existing Olympic record of 43.49 seconds, quite a feat!

Will the regression line's prediction be accurate? In the last two decades, there appears to be more of a cyclical (up and down) trend. Could winning times continue to drop at the same average rate? Extensive searches for talented potential athletes and improved full-time training methods can lead to decreased winning times, but ultimately, there will be a physical limit for humans.

Note that there were some unusual data points of 46.7 seconds in 1956 and 43.80 in 1968, which are far above and far below the regression line.

If we restrict ourselves to looking just at the most recent winning times, beyond 1968, for Olympic winning times in 1972 and beyond (10 winning times), we have the following scatterplot and regression line.



Using the most recent ten winning times, our regression line is $y = -0.025x + 93.834$.
When $x = 2012$, the prediction is $y = -0.025(2012) + 93.834 \approx 43.5$ seconds. This line predicts a winning time of 43.5 seconds for 2012 and that would indicate an excellent time close to the existing record of 43.49 seconds, but not dramatically below it.

Note too that for $r^2 = 0.5351$ and for the negatively sloping line, the correlation coefficient is $r = -\sqrt{0.5351} = -0.73$, not as strong as when we considered the time period going back to 1948. The most recent set of 10 winning times do not visually exhibit as strong a linear trend as the set of 16 winning times dating back to 1948.

CONCLUSION:

I have examined two linear models, using different subsets of the Olympic winning times for the men's 400 meter dash and both have moderately strong negative correlation coefficients. One model uses data extending back to 1948 and predicts a winning time of 43.1 seconds for the 2012 Olympics, and the other model uses data from the most recent 10 Olympic games and predicts 43.5 seconds. My guess is that 43.5 will be closer to the actual winning time. We will see what happens later this summer!

UPDATE: When the race was run in August, 2012, the winning time was 43.94 seconds.