ATLANTA PUBLIC SCHOOLS
Mathematics & Science Initiative
Making A Difference

Atlanta Public Schools

Teacher's Curriculum Supplement

Mathematics II: Unit 6
Finding the Best Model

GE Foundation

This document has been made possible by funding from the GE Foundation Developing Futures grant, in partnership with Atlanta Public Schools. It is derived from the Georgia Department of Education Math II Framework and includes contributions from Georgia teachers. It is intended to serve as a companion to the GA DOE Math II Framework Teacher Edition.

Preface

We are pleased to provide this supplement to the Georgia Department of Education's Mathematics II Framework. It has been written in the hope that it will assist teachers in the planning and delivery of the new curriculum, particularly in these first years of implementation. This document should be used with the following considerations.

- The importance of working the tasks used in these lessons cannot be overstated. In planning for the teaching of the Georgia Performance Standards in Mathematics teachers should work the tasks, read the teacher notes provided in the Georgia Department of Education's Mathematics II Framework Teacher Edition, and *then* examine the lessons provided here.

- This guide provides day-by-day lesson plans. While a detailed scope and sequence and established lessons may help in the implementation of a new and more rigorous curriculum, it is hoped that teachers will assess their students informally on an on-going basis and use the results of these assessments to determine (or modify) what happens in the classroom from one day to the next. Planning based on student need is much more effective than following a pre-determined timeline.

- It is important to remember that the Georgia Performance Standards provide a balance of concepts, skills, and problem solving. Although this document is primarily based on the tasks of the Framework, we have attempted to help teachers achieve this all important balance by embedding necessary skills in the lessons and including skills in specific or suggested homework assignments. The teachers and writers who developed these lessons, however, are not in your classrooms. It is incumbent upon the classroom teacher to assess the skill level of students on every topic addressed in the standards and provide the opportunities needed to master those skills.

- In most of the lesson templates, the sections labeled *Differentiated support/enrichment* have been left blank. This is a result of several factors, the most significant of which was time. It is hoped that as teachers use these lessons, they will contribute their own ideas, not only in the areas of differentiation and enrichment, but in other areas as well. Materials and resources abound that can be used to contribute to the teaching of the standards.

On the topic of differentiation, it is critical to reiterate that many of the strategies used in a standards-based mathematics classroom promote differentiation. These strategies include

- the use of rich tasks with multiple points of entry and more than one path to a solution,
- flexible grouping of students,
- multiple representations of mathematical concepts,
- writing in mathematics,
- monitoring of progress through on-going informal and formative assessments, and
- analysis of student work.

We hope that teachers will incorporate these strategies in each and every lesson.

It is hoped that you find this document useful as we strive to raise the mathematics achievement of all students in our district and state. Comments, questions, and suggestions for inclusions related to this document may be emailed to Dr. Dottie Whitlow, Executive Director, Mathematics and Science Department, Atlanta Public Schools, dwhitlow@atlantapublicschools.us.

This document has been made possible by funding from the GE Foundation Developing Futures grant, in partnership with Atlanta Public Schools. It is derived from the Georgia Department of Education Math II Frameworks and includes contributions from Georgia teachers. It is intended to serve as a companion to the GA DOE Math II Framework Teacher Edition.

Explanation of the Terms and Categories Used in the Lesson Template

Task: This section gives the suggested number of days needed to teach the concepts addressed in a task, the task name, and the problem numbers of the task as listed in the Georgia Department of Education's Mathematics II Framework Teacher Edition (GaDOE TE).

In some cases new tasks or activities have been developed. These activities have been named by the writers.

Standard(s): Although each task addresses many Math II standards and uses mathematics learned in earlier grades, in this section, only the key standards addressed in the lesson are listed.

New Vocabulary: Vocabulary is listed here the *first* time it is used. It is strongly recommended that teachers, particularly those teaching Math Support, use interactive word walls. Vocabulary listed in this section should be included on the word walls and previewed in Math Support.

Mathematical concepts/skills: Major concepts addressed in the lesson are listed in this section whether they are Math II concepts or were addressed in earlier grades or courses.

Prior knowledge: Prior knowledge includes only those topics studied in previous grades or courses. It does not include Math II content taught in previous lessons.

Essential Question(s): Essential questions may be daily and/or unit questions.

Suggested materials: This is an attempt to list all materials that will be needed for the lesson, including consumable items, such as graph paper; and tools, such as graphing calculators and compasses. This list does not include those items that should always be present in a standards-based mathematics classroom such as markers, chart paper, and rulers.

Warm-up: A suggested warm-up is included with every lesson. Warm-ups should be brief and should focus student thinking on the concepts that are to be addressed in the lesson.

Opening: Openings should set the stage for the mathematics to be done during the work time. The amount of class time used for an opening will vary from day-to-day but this should not be the longest part of the lesson.

Worktime: The problem numbers have been listed and the work that students are to do during the worktime has been described. A student version of the day's activity follows the lesson template in every case. In order to address all of the standards in Math II, some of the problems in some of the original GaDOE tasks have been omitted and less time consuming activities have been substituted for those problems. In many instances, in the student versions of the tasks, the writing of the original tasks has been simplified. In order to preserve all vocabulary, content, and meaning it is important that teachers work the original tasks as well as the student versions included here.

Teachers are expected to both facilitate and provide some direct instruction, when necessary, during the work time. Suggestions related to student misconceptions, difficult concepts, and deeper meaning have been included in this section. However, the teacher notes in the GaDOE Math II Framework are exceptional. In most cases, there is no need to repeat the information provided there. Again, it is imperative that teachers work the tasks and read the teacher notes that are provided in GaDOE support materials.

Questioning is extremely important in every part of a standards-based lesson. We included suggestions for questions in some cases but did not focus on providing good questions as extensively as we would have liked. Developing good questions related to a specific lesson should be a focus of collaborative planning time.

Closing: The closing may be the most important part of the lesson. This is where the mathematics is formalized. Even when a lesson must be continued to the next day, teachers should stop, leaving enough time to "close", summarizing and formalizing what students have done to that point. As much as possible students should assist in presenting the mathematics learned in the lesson. The teacher notes are all important in determining what mathematics should be included in the closing.

Homework: In some cases, homework suggestions are provided. Teachers should use their resources, including the textbook, to assign homework that addresses the needs of their students.

Homework should be focused on the skills and concepts presented in class, relatively short (30 to 45 minutes), and include a balance of skills and thought-provoking problems.

Differentiated support/enrichment: On the topic of differentiation, it is critical to reiterate that many of the strategies used in a standards-based mathematics classroom promote differentiation. These strategies include

- the use of rich tasks with multiple points of entry and more than one path to a solution
- flexible grouping of students
- multiple representations of mathematical concepts
- writing in mathematics
- monitoring of progress through on-going informal and formative assessments
- and analysis of student work.

Check for understanding: A check for understanding is a short, focused assessment-a ticket out the door, for example. There are many good resources for these items, including the GaDOE culminating task at the end of each unit and the *Mathematics II End-of-Course Study Guide*. Both resources can be found on-line at www.georgiastandards.org, along with other GaDOE materials related to the standards. Problem numbers from the GaDOE culminating task have been listed with the appropriate lessons in this document.

Resources/materials for Math Support: Again, in some cases, we have provided materials and/or suggestions for Math Support. This section should be personalized to your students, class, and/or school, based on your resources.

Table of Contents

Mathematics II
Unit 6

☺☺

Unit 1 Timeline

Task 1: Oh….the Butterfly Ballot!	3 days
Task 2: Straight of Curved?	1 day
Task 3: Orbital Debris	1 day
Task 4: Causation versus Correlation	1 day

Task Notes

The importance of working the tasks used in these lessons cannot be overstated. In planning for the teaching of the Georgia Performance Standards in Mathematics, teachers should work the Student Tasks, read any corresponding teacher notes provided in the Georgia Department of Education's *Mathematics II Framework Teacher Edition*, and *then* examine the lessons provided here.

The tasks provided in this Supplement are based on the content of Unit 6 of the Georgia Department of Education's *Mathematics II Framework*. We suggest, as always, that teachers use this Supplement along with the GaDOE Teacher Edition and the *Mathematics II End-of-Course Study Guide* which can be found on-line at www.georgiastandards.org.

Task 1: Oh….the Butterfly Ballot!
The concepts and skills addressed in this task include:

- determining whether a linear association exists between two quantities that vary;
- measuring the strength of a linear association using the Pearson Correlation Coefficient;
- finding good fits to data using visual inspection ("eyeballing"), median-median lines and least squares regression lines;
- using linear models to predict values;
- comparing lines of fit;
- understanding properties of the least squares regression line;
- understanding the effect of outliers on regression models;
- using appropriate technology to conduct statistical analysis; and,
- determining what questions can be answered and what conclusions can be drawn from a statistical analysis.

This task was adapted from *Statistics: The Art and Science of Learning from Data* by Agresti and Franklin. It does not appear in the GaDOE Mathematics II Framework. It addresses the content covered in the first GaDOE task of Unit 6, *Traveling on the Turnpike*, with the exception of writing absolute value functions as piecewise functions and solving absolute value equations. Both of these topics were addressed in Unit 5 of the Supplement.

Teacher notes for this task begin on page 26.

Task 2: Straight or Curved
The concepts and skills addressed in this task include:

- determining whether an association exists between two quantities that vary;
- measuring the strength of a linear association using the Pearson Correlation Coefficient;
- using visual inspection and the sum of squares of the residuals to determine which regression curve, linear or quadratic, better approximates the relationship between two variables represented by a set of data points;
- using regression equations to make predictions; and,
- using appropriate technology to conduct statistical analyses.

This task does not appear in the GaDOE Mathematics II Framework. Teacher notes begin on page 45.

.

Task 3: Orbital Debris
The concepts and skills addressed in this task include:

- determining whether an association exists between two quantities that vary;
- measuring the strength of a linear association using the Pearson Correlation Coefficient;
- using visual inspection and the sum of squares of the residuals to determine which regression curve, linear or quadratic, better approximates the relationship between two variables represented by a set of data points;
- using regression equations to find interpolated values;
- examining the feasibility of extrapolation using regression equations; and,
- using appropriate technology to conduct statistical analyses.

All items of the GaDOE task are addressed. See notes beginning on page 51 of the GaDOE *Mathematics II Framework Teachers' Edition*, Unit 6.

Task 4: Causation versus Correlation
The concepts and skills addressed in this task include:

- understanding the difference between correlation and causation, and
- realizing that confusion between correlation and causation is pervasive in popular media.

All items of the GaDOE task are addressed. See notes beginning on page 65 of the GaDOE *Mathematics II Framework Teachers' Edition*, Unit 6.

ATLANTA PUBLIC SCHOOLS
Mathematics & Science Initiative
Making A Difference

Atlanta Public Schools

Teacher's Curriculum Supplement

Mathematics II: Unit 6

Task 1: Oh….the Butterfly Ballot!

Mathematics II

Task 1: *Oh….the Butterfly Ballot!* **Day 1/3**

(adapted from *Statistics: The Art and Science of Learning from Data* by Agresti and Franklin.)

See teacher notes beginning on page 26.

MM2D2. Students will determine an algebraic model to quantify the association between two quantitative variables.

a. Gather and plot data that can be modeled with linear and quadratic functions.
b. Examine the issues of curve fitting by finding good linear fits to data using simple methods such as the median-median line and "eyeballing".
c. Understand and apply the process of linear and quadratic regression for curve fitting using appropriate technology.

New vocabulary: response variable, explanatory variable, association, Pearson correlation coefficient, "eyeball" line

Mathematical concepts/skills:

- posing questions that can be answered by data
- determining whether there is an association between two variables
- examining the strength of the linear relationship between two variables as measured by the Pearson Correlation Coefficient
- using "eyeballing" or visual approximation to determine a line of best fit for a set of data

Prior knowledge:

- plotting sets of data
- using graphs, specifically scatterplots, to examine sets of data
- determining an equation of a line given two points or the slope and one point
- using "eyeballing" or visual approximation to determine a line of best fit for a set of data

Essential question(s): How can I determine whether there is an association between two quantities that vary? What information can I obtain from a model used to predict the value of a response variable, given the value of the explanatory variable?

Suggested materials: graph paper, colored pencils, graphing calculators, raw spaghetti, rulers, Smartview and computer with projection device (if possible)

Warm-up: Ask students to read the introduction to the task and examine the picture of the ballot.

Opening: Have a discussion of the issues.

- You might ask students whether they find the ballot confusing.
- Some students may be familiar with the demographics of Palm Beach County and realize that many residents of the county are older (more than 22% of the population is over 65 years of age compared to 12.8% nationally). Students may bring up eyesight and other factors.

Before beginning the task, students are asked to discuss the two questions below with their groups or as a class. However you choose to have the discussion, be sure that students are respectful of each others' opinions.

- *Why do you think we might compare these two sets of data (votes for Perot in 1996 and votes for Buchanan in 2000) to explore whether the Buchanan vote in Palm Beach County was unusually high?*
- *Do you think using this comparison is a reasonable way to explore the question? Why or why not?*

Note: There are actually 67 counties in Florida. Due to time constraints, we decided to examine the vote in only 25 counties. Interested students may want to see what happens when they examine the relationship between the votes for Perot and the votes for Buchanan in all 67 counties.

Worktime: Students should work in pairs or groups to complete *Problems 1 -6a.*
Teacher notes are provided beginning on page 26.

Because the plotting of 25 data points takes a significant amount of time, a plot of the data has been included immediately following this task. Teachers may choose to have students plot the data or give them the plot. In either case, students should answer *Problems 1* and *2*. Monitor work carefully to be sure that students understand which is the response variable and which is the explanatory variable.
Note: Whether students plot the data or are given the plot, it is important that each student have his/her own plot, determine their own regression line, and draw the line onto the plot.

When students have had ample time to complete *Problems 1 – 4*, have a whole class discussion of all 4 items. (See teacher notes.) After this discussion, allow students to investigate the strength of linear relationships using the applet at the website given in Problem 5.

Closing: Allow students to share the equations and graphs of their lines of fit. As they share, students should describe the methods they used for choosing their particular lines. Guiding questions might include:

- How did you choose your line?
- How did you find the equation of your line?
- Did you consider the point significantly removed from the other data in determining your line?
- How does your line compare to…. (another student who has shared)?

Homework:

Differentiated support/enrichment:

Check for Understanding:

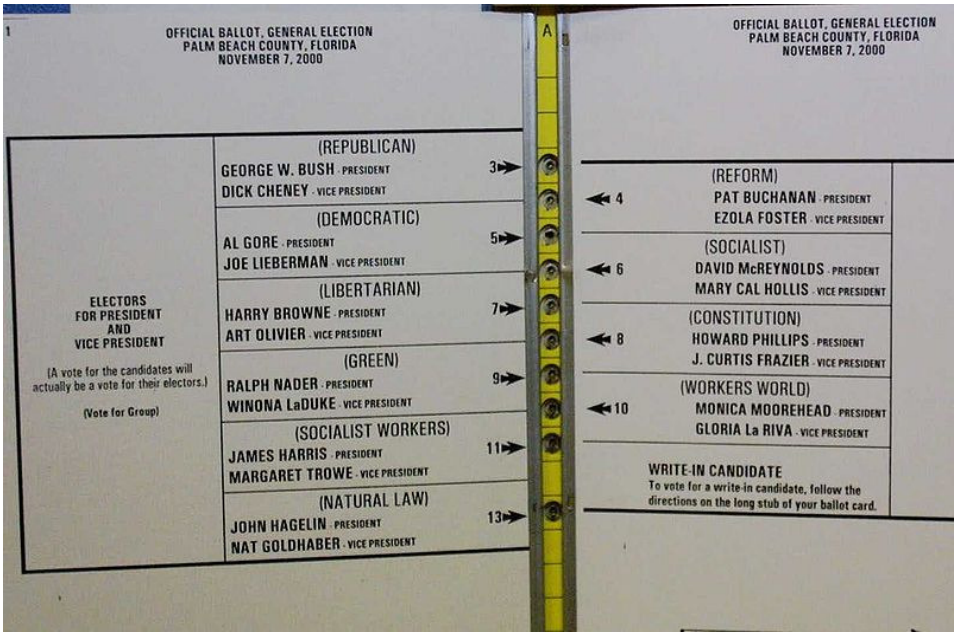
Resources/materials for Math Support: In addition to the new vocabulary listed above, students should preview:

- independent versus dependent variables
- plotting sets of data
- using graphs, specifically scatterplots, to examine sets of data
- determining an equation of a line given two points or the slope and one point
- using "eyeballing" or visual approximation to determine a line of best fit for a set of data

Mathematics II
Task 1: ***Oh….the Butterfly Ballot!***
Day 1 Student Task

The 2000 Presidential election was one of the most controversial in the history of our county. It was the closest election since 1876 and only the fourth election in which the electoral vote did not reflect the popular vote. The Democratic candidate was Al Gore and the Republican candidate was George W. Bush. Much of the controversy surrounded the awarding of Florida's 25 electoral votes to Bush.

Among the issues related to the Florida vote, was the use of the butterfly ballot in Palm Beach County. Initial election returns reported 3407 votes for the Reform Party candidate, Pat Buchanan. Some political analyst thought this total was surprisingly high. Many believed that most of these votes were actually intended for Gore but were wrongly cast for Buchanan because some voters found the ballot confusing. On the butterfly ballot, Bush appeared first in the left column, followed by Buchanan in the right column, and Gore in the left column. A picture of the ballot is shown below.



This task is designed to help explore whether the Buchanan vote in Palm Beach County in 2000 was, in fact, surprisingly high. In order to investigate this question, we will compare the votes cast in 25 Florida Counties for the 1996 Reform Party candidate, Ross Perot, to the votes cast for Buchanan, the Reform Party candidate in 2000.

Before you begin *Problem 1*, discuss the following two questions with your group or your class. And remember….everyone is allowed to have an opinion.

- Why do you think we might compare these two sets of data (votes for Perot in 1996 and votes for Buchanan in 2000) to explore whether the Buchanan vote in Palm Beach County was unusually high?
- Do you think using this comparison is a reasonable way to explore the question? Why or why not?

1. Statisticians often use different terminology from that of mathematicians. Read the following definitions and identify the mathematical terminology that would be used to describe each of the three terms.
   - The **response variable** is the outcome variable on which comparisons are made.
   - The **explanatory variable** defines the groups to be compared with respect to values on the response variable.
   - An **association** exists if the likelihood of a particular value for one variable depends on the value of the other variable.

2. Sketch a scatterplot of the data (or use the plot provided by your teacher). Be sure to consider the following questions:
   - Which variable should be considered the **explanatory variable** in this situation?
   - Which variable should be considered the **response variable**?
   - What scales are appropriate for the axes?

3. Examine your plot. Does there appear to be any **association** between the number of votes for Perot and the number of votes for Buchanan? If so, how would you describe the association?

4. When data points roughly follow a straight line trend, the variables are described as having a linear relationship. Sometimes the data points fall close to a line but more often there is considerable variability of the points about the line. A summary measure, describing the strength of the linear relationship between two variables was derived in 1896 by Karl Pearson, a British statistician. The **Pearson Correlation Coefficient** (denoted by r) takes values between -1 and 1, inclusive. The closer the absolute value of r is to 1, the closer the data points are to a straight line.

   What do you think a correlation of 1 might mean?
   What do you think a correlation of -1 might mean?
   What would a correlation of 0 indicate?

   Look again at your plot.
   Does there seem to be a positive association between the variables, a negative association, or no clear evidence of association? Explain your thinking.

   Can the trend be approximated reasonably well by a straight line? If so, do you think the points would be close to the line or would they scatter quite a bit?

   Are any of the observations unusual? If so, which ones? Why do you consider these observations unusual?

5. Use the correlation applet at http://www.stat.tamu.edu/~west/ph/coreye.html (or Google *Correlation by Eye*) to get a sense of the correlations for various scatterplots. Once you have used the applet to examine the correlations for several different plots, make a guess for the strength of the linear associatin between the variables in your scatterplot.

6. Three different methods of fitting lines to data are investigated in Math II: eyeballing, determining median-median lines, and the least squares regression method.
   - **Eyeballing** is simply approximating the equation of a line of best fit based on a visual inspection of a scatterplot of the data.
   - The **median-median line** is determined by using the medians of groups of the data.
   - The **least squares regression line** is the line that minimizes the sum of the squares of the vertical distances between the line and the points in the data set.

   a. Using raw spaghetti or a straight-edge, determine, by "eyeballing", an equation of a line of fit for your scatterplot. Graph your equation on the scatterplot using a colored pencil.

Mathematics II
Task 1: *Oh….the Butterfly Ballot* **Day 2/3**
(See teacher notes beginning on page 28.)

MM2D2. Students will determine an algebraic model to quantify the association between two quantitative variables.

a. Gather and plot data that can be modeled with linear and quadratic functions.
b. Examine the issues of curve fitting by finding good linear fits to data using simple methods such as the median-median line and "eyeballing".
c. Understand and apply the process of linear and quadratic regression for curve fitting using appropriate technology.

New vocabulary: median-median line, least squares regression line

Mathematical concepts/skills:

- posing questions that can be answered by data
- determining whether there is an association between two variables
- examining the strength of the linear relationship between two variables as measured by the Pearson Correlation Coefficient
- determining median-median lines and least squares regression lines to model the linear relationship between varying quantities
- using models to make predictions
- using appropriate technology to conduct statistical analyses

Prior knowledge:

- plotting sets of data
- using graphs, specifically scatterplots, to examine sets of data
- finding medians of sets of data
- determining an equation of a line given two points or the slope and one point
- using "eyeballing" or visual approximation to determine a line of best fit for a set of data
- determining the distance between a point and a line

Essential question(s): How can I determine whether there is an association between two quantities that vary? What information can I obtain from a model used to predict the value of a response variable, given the value of the explanatory variable? How can technology help me conduct statistical analyses?

Suggested materials: graph paper, colored pencils, graphing calculators, rulers, Smartview or projection devices for technology being used (if possible)

Warm-up: Post the following:

Find the vertical distance between the line $y = 2x - 5$ and the point (3, 7).

Opening: We have chosen to open this lesson by "walking" students through the process of calculating a median-median line using 5 points. (Steps for this calculation are included immediately following the teacher notes for this lesson). It is important for students to understand this process. It is even more important to ultimately understand that, because the median-median line is calculated based on medians of the data, it is more resistant to outliers than the least squares regression line which is calculated using means. It is **not** important for students to calculate a median-median line using large sets of data. Once the process has been explained and is understood, students should use appropriate technology to find these lines.

Begin by showing students how to find the median-median line for the following five data points: (1, 2), (3, 1), (4, 4), (5, 2), (7, 3).

1. *Separate the data into three groups of equal size according to the values of the horizontal coordinate. If the number of data points is not a multiple of three, choose the outer groups first and allow the middle group to contain one more or one less point.*

Group 1	Group 2	Group 3
(1, 2) (3, 1)	(4,4)	(5, 2) (7, 3)

2. *Find a summary point for each group based on the median x-value and the median y-value of the points in that group. Label the summary points* (x_1, y_1), (x_2, y_2), *and* (x_3, y_3), *respectively.*

Group 1	Group 2	Group 3
$(x_1, y_1) = (2, 1.5)$	$(x_2, y_2) = (4,4)$	$(x_3, y_3) = (6, 2.5)$

3. *Find the equation of the line L through the points* (x_1, y_1) *and* (x_3, y_3), *the summary points of the outer groups.*

$$\text{Line } L: \quad y = \frac{1}{4}x + \frac{7}{4}$$

4. *The median-median line is the line parallel to L, one-third of the vertical distance from L to the middle summary point* (x_2, y_2). *Translate L by doing the following:*
   a. *Find the y-coordinate of the point on L with the same x-coordinate as the middle summary point.*

$$\text{On line } L, \text{ when } x = 4, \quad y = \frac{11}{4}$$

   b. *Find the vertical distance between the middle summary point and the line by subtracting these y-values.*

$$4 - \frac{11}{4} = \frac{5}{4}$$

c. *Translate the line L up or down (toward the middle summary point), one-third of the distance from the line L to the middle summary point.*

$$\frac{1}{3} \bullet \frac{5}{4} = \frac{5}{12}$$

The median-median line for this set of data is:

$$y = \frac{1}{4}x + \frac{7}{4} + \frac{5}{12} \quad or \quad y = \frac{1}{4}x + \frac{13}{6}$$

Worktime: Students should work in pairs or groups to complete *Problems 6b – 7.*
Note: Teacher notes for these problems begin on page 29.

The amount of time needed to complete these items will depend on how much your students already know about using graphing calculators or the statistical software chosen to conduct these analyses.

If using calculators, by the end of this lesson students should be able to:

- enter and edit lists;
- calculate regression curves and automatically enter them into the *Y=* menu;
- calculate *y* values using the *Y-VARS* , *Function* menu; and
- graph regression lines on scatterplots of data.

Note: A set of calculator instructions is included on page 36.

Closing: Discuss *Problem 7*, thoroughly. You may want to choose one group of students to facilitate the discussion of each part of the problem.

Homework:

Differentiated support/enrichment:

Check for Understanding:

Resources/materials for Math Support: In addition to the new vocabulary listed above, students should preview:

- using graphs, specifically scatterplots, to examine sets of data
- finding medians of sets of data
- determining the distance between a point and a line

Mathematics II
Task 1: *Oh….the Butterfly Ballot*
Day 2 Student Task

6. Three different methods of fitting lines to data are investigated in Math II: eyeballing, determining median-median lines, and the least squares regression method.
   - **Eyeballing** is simply approximating the equation of a line of best fit based on a visual inspection of a scatterplot of the data.
   - The **median-median line** is determined by using the medians of groups of the data.
   - The **least squares regression line** is the line that minimizes the sum of the squares of the vertical distances between the line and the points in the data set.

   b. Use the instructions provided by your teacher to determine an equation of a median-median line for your data. Using a different color from the one you used in *part a*, draw the median-median line on your plot.

   c. Use your calculator to determine the least squares regression line for your plot. What is the correlation coefficient for this line? Graph the line using a third color.

7. Consider the three lines of fit that you have drawn on your scatterplot.
   a. In the case of each line, what does the slope represent? What does the *y*-intercept represent?

   b. How are the lines alike? How are they different?

   c. Use each of your lines to predict the vote for Buchanan in the counties given in the table below. Use the table to organize your information.

	Predicted Vote for Buchanan			
	Liberty	Manatee	Hillsborough	Palm Beach
Eyeballing				
Median-Median Line				
Least Squares Regression Line				

   d. Which line, in your opinion appears to best fit the data? Explain your thinking.

Mathematics II
Task 1: *Oh….the Butterfly Ballot* **Day 3/3**
(See teacher notes beginning on page 31.)

MM2D2. Students will determine an algebraic model to quantify the association between two quantitative variables.

a. Gather and plot data that can be modeled with linear and quadratic functions.
b. Examine the issues of curve fitting by finding good linear fits to data using simple methods such as the median-median line and "eyeballing".
c. Understand and apply the process of linear and quadratic regression for curve fitting using appropriate technology.

New vocabulary: residual, prediction error, the sum of squares of the residuals, regression outlier, influential outlier

Mathematical concepts/skills:

- finding the residuals (prediction error) for points in a data set based on the least squares regression line
- understanding the properties of the least squares regression line
- comparing lines of fit based on the sum of the squares of the residuals
- determining whether a data point is a regression outlier and whether a regression outlier is also an influential outlier
- determining what questions can be answered and what conclusions can be drawn from a regression analysis

Prior knowledge:

- determining the distance between a point and a line

Essential question(s): What are residuals and how do they help me determine how well a function models a given set of data? What effect do outliers have on regression models and how can regression models help me answer questions related to those outliers? What conclusions can I reasonably draw from a statistical analysis of a data set?

Suggested materials: graph paper, colored pencils, graphing calculators, rulers, Smartview or projection devices for technology being used (if possible)

Warm-up: Post the following:

Use your scatterplot, the least squares regression line determined in Item 6c, and the table completed in Item 7c to help you answer the following questions:

What is the vertical distance between the least squares regression line and the data point on your scatterplot representing the votes for Perot and Buchanan in Liberty County? What does this distance tell you?

Opening: Help students begin today's assignment by discussing the warm-up. Introduce the formal vocabulary related to residuals by having students read the introduction to *Problem 8* and then asking several students to summarize what they have read.

Worktime: Students should complete *Problems 8 – 14* of the task. *Problem 14* asks students to write a short paper to summarize their work and discuss any conclusions that they might draw from this analysis. This would be an excellent homework assignment and/or assessment of the understandings they have gained.
Note: Teacher notes begin on page 31.

Problem 8 is designed to help students gain an understanding of residuals using a concrete, visual approach. Problem 9 guides them to generalize this understanding to algebraic expressions that can be used to calculate residuals and the sums of squares of residuals. This symbolic generalization, in turn, should help them understand plots of residuals and the calculator operations used to find the sums of squares of residuals.

Note: A set of calculator instructions is included on page 36.

Closing: Discuss *Problem 12*, thoroughly. (See teacher notes.)

Homework:

Differentiated support/enrichment:

Check for Understanding:

Resources/materials for Math Support: Students should preview the new vocabulary listed above. Students will probably need extra help using graphing calculators, particularly to find sums of squares of residuals.

Mathematics II
Task 1: ***Oh….the Butterfly Ballot***
Day 3 Student Task

8. The **residual** for any point in a data set is the **prediction error** that occurs when a regression line is used to predict the *y* value for a given value of *x*. If the data point lies above a regression line, its **residual** is defined to be the vertical distance between the line and the point. If the data point lies below the regression line, its residual is the opposite of the vertical distance from the point to the line. (Some refer to this as the *signed* distance from the line since the residual is positive when the data point is above the line and negative when the data point is below the line).

    a. On your scatterplot, label each of the 4 data points representing the counties in the table below with their ordered pairs.

    b. Look again at the **least squares regression line** that you found in *Item 6c*. Mark the points on your least squares line that correspond to the counties in the table. For each point, draw a line segment that represents the vertical distance from the regression line to the corresponding data value in the scatterplot.

    c. Find the residuals related to the least squares regression line for the points representing the four counties in the table.

Residuals			
Liberty	Manatee	Hillsborough	Palm Beach

    d. Explain what each of these residuals tells you?

    e. Explain how would you compute the residual for any point in the data set?

The least squares regression line is often represented as $\hat{y} = ax + b$, where "*y*-hat" is the predicted value of *y* for a given *x*. As a result of the formula used to calculate the line, the following properties hold for the least squares line of any data set:

- The sum of the residuals is 0.
- The regression line includes the point $(\overline{x}, \overline{y})$.
- The regression line minimizes the sum of the squares of the residuals.
- There is a *unique* least squares regression line for a given set of data.

9. Consider that a data set has n distinct data points.
    a. Write an ordered pair for the i^{th} data point.

    b. Using the symbol for "*y*-hat", write an ordered pair for the point on the least squares regression line corresponding to the i^{th} data point.

c. Using the ordered pairs from *parts a* and *b*, write an algebraic expression for the residual of the i^{th} data point.

d. In *Item 6*, the least squares regression line is described as the line that minimizes the sum of the squares of the vertical distances between the line and the points in the data set. In other words, the least squares regression line minimizes the sum of the squares of the residuals. Write an algebraic expression for the sum of the squares of the residuals of a set of data with n distinct data points.

10. Use your calculator to compare the sum of squares of the residuals for the median-median line and the least squares regression line determined in *Item 6*. What did you find? Which line do you think appears to be the better fit for this data? Why do you think this might be true?

11. A **regression outlier** is defined as a point that is well removed from the trend of the other data. Are there regression outliers in our data? If so, what are the outliers?

12. An **influential outlier** is a regression outlier with an *x*-value that is relatively high or low compared to the rest of the data. Because the derivation of a least squares regression line is based on means of the data, influential outliers have a large effect on this particular regression line.

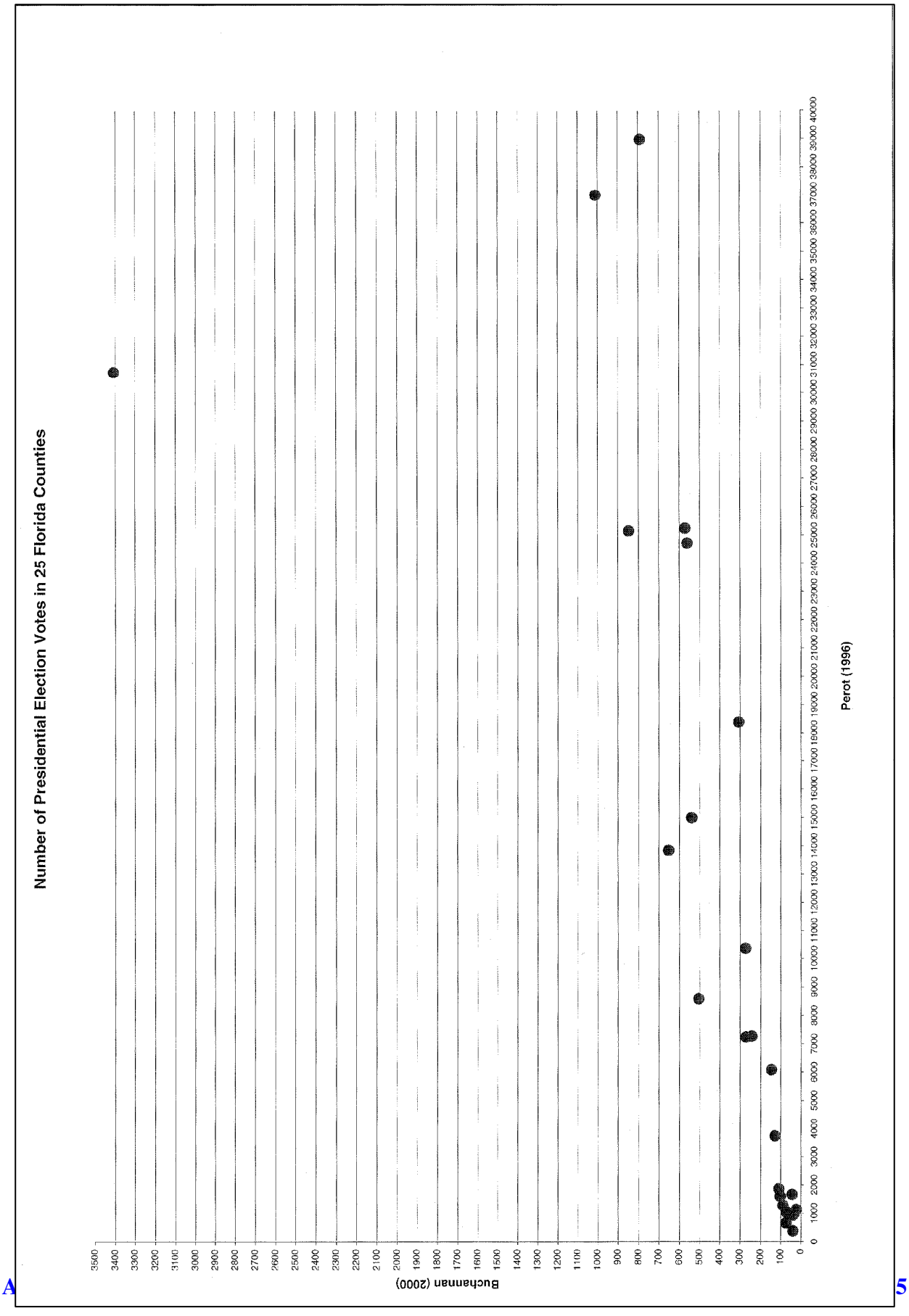
    Are there influential outliers in your data? If so, recalculate the regression line without the influential outliers and then recalculate the sum of squares of the residuals for your new regression equation. Compare this value with the sum of squares of the residuals for the least squares regression line in *Item 10*. What does this comparison tell you?

13. Use your *new* regression line to predict the votes for Buchanan in Liberty, Manatee, Hillsborough, and Palm Beach counties.

14. As a statistician, what conclusions might you draw from the analyses you have conducted in this task? Write a short paper to summarize your work and explain your thinking.

Numbers of Votes in 25 Florida Counties' Presidential Election

County	Perot (1996)	Buchanan (2000)	Gore (2000)	Bush (2000)
Baker	667	73	2392	5610
Brevard	25249	570	97318	115185
Broward	38964	789	386518	177279
Citrus	7244	270	25525	29766
Desoto	965	36	3322	4256
Duval	13844	652	107864	152098
Escambia	8587	504	40958	73029
Gadsden	938	39	9565	4750
Gulf	1054	71	2397	3550
Hendry	1135	22	3240	4747
Hernando	7272	242	32644	30646
Highlands	3739	127	14167	20206
Hillsborough	25154	847	169557	180760
Jackson	1602	102	6868	9138
Lee	18389	305	73560	106141
Liberty	376	39	1011	1316
Manatee	10360	272	49169	57948
Miami Dade	24722	560	328764	289492
Okeechobee	1666	43	4588	5057
Osceola	6091	145	28181	26212
Palm Beach	30739	3407	268945	152846
Pinellas	36990	1010	200212	184884
Polk	14991	538	74977	90101
Suwannee	1874	108	4075	8006
Washington	1287	88	2796	4983


Number of Presidential Election Votes in 25 Florida Counties
Buchanan (2000)
Perot (1996)
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 15000 16000 17000 18000 19000 20000 21000 22000 23000 24000 25000 26000 27000 28000 29000 30000 31000 32000 33000 34000 35000 36000 37000 38000 39000 40000
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500

Teacher Notes
Task 1: *Oh….the Butterfly Ballot!*

This task is adapted from *Statistics: The Art and Science of Learning from Data* by Agresti and Franklin. Additional information on the topic can be found on pages 102 and 103 of the text.

Teachers should open the lesson by allowing students to read the introduction to the task and examine the picture of the ballot. Have a discussion of the issues.

- You might ask students whether they find the ballot confusing.
- Some students may be familiar with the demographics of Palm Beach County and realize that many residents of the county are older (more than 22% of the population is over 65 years of age compared to 12.8% nationally). Students may bring up eyesight and other factors.

Before beginning the task, students are asked to discuss the two questions below with their groups or as a class. However you choose to have the discussion, be sure that students are respectful of each others' opinions.

- Why do you think we might compare these two sets of data (votes for Perot in 1996 and votes for Buchanan in 2000) to explore whether the Buchanan vote in Palm Beach County was surprisingly high?
  Students need to realize that we are comparing the votes cast for Perot in 1996 to the votes cast for Buchanan in 2000. The idea is that, in a given county, there may be a relationship between the number of people who voted for a Reform Party candidate in 1996 and the number of people who voted for a Reform Party candidate in 2000. If such a relationship exists, we might use it to predict a likely number of votes for Buchanan in Palm Beach County. Comparing this prediction to the actual number of votes may shed some light on our question of whether this vote was unusually high.

- Do you think using this comparison is a reasonable way to explore the question? Why or why not?
  Students are very likely to interject their personal biases in considering whether the comparison made in the task is reasonable. Honor all opinions. Ultimately, it will be important for students to realize that we are not trying to establish causation or to examine mitigating circumstances surrounding the ballot. In this task, we are simply trying to establish whether there is an **association** between the number of votes cast for Perot and the number of votes cast for Buchanan. If there is such an association, what might we learn from examining a prediction based on a function approximating this association? We will leave it to the students (and the pundits) to draw other conclusions.

1. Statisticians often use different terminology from that of mathematicians. Read the following definitions and identify the mathematical terminology that would be used to describe each of the three terms.
   - The **response variable** is the outcome variable on which comparisons are made.
   - The **explanatory variable** defines the groups to be compared with respect to values on the response variable.
   - An **association** exists if the likelihood of a particular value for one variable depends on the value of the other variable.

Using mathematical terminology, the response variable would be referred to as the dependent variable, the explanatory variable is called the independent variable, and an association is referred to as a relation.

2. Sketch a scatterplot of the data (or use the plot provided by your teacher). Be sure to consider the following questions:
   Students may draw their own plots or, as a time saving measure, you may choose to give them the plot. (We have included a scatterplot of the data in this teacher supplement on page 25). In either case, the following questions should be answered.
   - Which variable should be considered the **explanatory variable** in this situation?
     The number of votes for Perot
   - Which variable should be considered the **response variable**?
     The number of votes for Buchanan
   - What scales are appropriate for the axes?
     Students may use different scales. We used a scale of 1000 for the explanatory variable and 100 for the response variable.

3. Examine your plot. Does there appear to be any **association** between the number of votes for Perot and the number of votes for Buchanan? If so, how would you describe the association?
   Allow students to describe any association they see using informal language. They may make statements such as the ones listed below.
   - There is a positive association between the number of votes for Perot and the number of votes for Buchanan.
   - As the number of votes for Perot increases, the number of votes for Buchanan generally increases.
   - The relationship between the number of votes for Perot and the number of votes for Buchanan appears to be linear.

4. When data points roughly follow a straight line trend, the variables are described as having a linear relationship. Sometimes the data points fall close to a line but more often there is considerable variability of the points about the line. A summary measure, describing the strength of the linear relationship between two variables was derived in 1896 by Karl Pearson, a British statistician. The **Pearson Correlation Coefficient** (denoted by r) takes values between -1 and 1, inclusive. The closer the absolute value of r is to 1, the closer the data points are to a straight line.

   What do you think a correlation of 1 might mean?
   Again, allow student language at this point to be informal. By the end of this task, students should know that a correlation of 1 or -1 means that the data in a set is perfectly linear. If a regression line is used to predict a y value for any given x, the prediction can be made with 100% accuracy. A correlation of positive 1 means that there is a positive association between the variables (as the value of the explanatory variable increases, the value of the response variable increases) and that the slope of the regression line is positive.

What do you think a correlation of -1 might mean?
A correlation of -1 means that there is a negative association between the variables (as the value of the explanatory variable increases, the value of the response variable decreases) and the slope of the regression line is negative.
What would a correlation of 0 indicate?
A correlation of 0 would indicate that there is no clear linear association between the variables.

Look again at your plot.Does there seem to be a positive association between the variables, a negative association, or no clear evidence of association? Explain your thinking.

Can the trend be approximated reasonably well by a straight line? If so, do you think the points would be close to the line or would they scatter quite a bit?

Are any of the observations unusual? If so, which ones?
As a result of work done fitting lines to data in Grade 8, most students should see that there appears to be a positive association between the number of votes for Perot and the number of votes for Buchanan. Generally, as the number of votes for Perot increases, the number of votes for Buchanan increases. The data can be approximated fairly well by a linear function, although the points will be scattered about the line as opposed to lying very close no matter what line is used.
The point (30,739, 3407) is significantly removed from the rest of the data.

5. Use the correlation applet at http://www.stat.tamu.edu/~west/ph/coreye.html (or Google *Correlation by Eye*) to get a sense of the correlations for various scatterplots. Once you have used the applet to examine the correlations for several different plots, make a guess for the strength of the linear correlation between the variables in your scatterplot.
Practice using this applet is an engaging way to give students a frame of reference for the strength of the linear correlation for a set of data. Students may practice individually, if they have access to a computer, or the applet may be used as a whole-class activity. One way to do this is to have students write their guesses for *r* on a sheet of paper and then award a small prize to the student whose guess is closest to *r*.

   The strength of the linear association for the 25 data points representing votes for Perot in 1996 and votes for Buchanan in 2000 is .67. This value of *r* is based on the least squares regression line calculated using the TI-84 calculator. The Palm Beach outlier has a huge influence on the least squares line. Students who try to consider the outlier in determining *r* will probably have a lower guess that those who choose to ignore the outlier. Later on in the task, students will have the opportunity to check their guesses by determining least squares regression lines with and without the outlier.

6. Three different methods of fitting lines to data are investigated in Math II: eyeballing, determining median-median lines, and the least squares regression method.
   - **Eyeballing** is simply approximating the equation of a line of best fit based on a visual inspection of a scatterplot of the data.

- The **median-median line** is determined by using the medians of groups of the data.
- The **least squares regression line** is the line that minimizes the sum of the squares of the vertical distances between the line and the points in the data set.

a. Using raw spaghetti or a straight-edge, determine, by "eyeballing", an equation of a line of fit for your scatterplot. Graph your equation on the scatterplot using a colored pencil.
Answers will vary. Having vetted this task with several different groups, it seems that when visually approximating a line of fit, more people than not basically ignored the outlier. Slopes of the "eyeball" lines were generally smaller than those for the median-median line and certainly smaller than the slope of the least squares regression line. A sample equation of a line is given here but there is no one correct answer: $y = .023x + 86$

b. Use the instructions provided by your teacher to determine an equation of a median-median line for your data. Using a different color from the one you used in *part a*, draw the median-median line on your plot.
Important note: For purposes of this Supplement, we have chosen to open this lesson by "walking" students through the process of calculating a median-median line using 5 points. (Steps for this calculation are included immediately following these teacher notes). It is important for students to understand this process. It is even more important to understand that, because the median-median line is calculated based on medians of the data, it is more resistant to outliers than the least squares regression line which is calculated using means. It is **not** important for students to calculate a median-median line using large sets of data. Once the process has been explained and is understood, students should use appropriate technology to find these lines.

Calculator notes: When performing regression analysis using the TI calculator, it is important that students have Diagnostics turned On. To turn diagnostics on, press 2^{nd}, 0 to obtain the catalog. Toggle down (or press D, which is above the x^{-1} key and then toggle down) to *DiagnosticOn* and press enter twice or until the calculator screen shows *Done*. To find the median-median line students will need to enter the *x*-values of the data set into one list and the *y*-values into another list. The Med-Med command is number 3 on the STAT , CALC menu. Students can calculate the median-median line and automatically input it as an equation in the *Y*= menu by entering *Med-Med* (*L# of x-values, L# of y-values, Y-variable*). For example, if the *x*-values are stored in List 2 and the *y*-values are stored in List 3 and you would like the equation for the median-median line stored in Y_2 = , the entry would read Med-Med (*L2, L3,* Y_2). Y_2 can be obtained from the VARS menu. If the *x*-values for a set of data are stored in *L1* and the *y*-values are stored in *L2*, entering *Med-Med and the Y variable* is sufficient because the calculator defaults to *L1* for *x*-values and *L2* for *y*-values in the absence of other list numbers.

The median-median line generated by the TI-84 calculator for the data in this task, with the slope and *y*-intercept rounded to 4 decimal places, is $y = .0258x + 37.6273$. Students should generate this line using the calculator and then draw it on their scatterplot using a color different from the one used to draw their "eyeball" line.

c. Use your calculator to determine the least squares regression line for your plot. What is the correlation coefficient for this line? Graph the line using a third color.
The least squares regression line should be determined using appropriate technology. The theory and calculations of the slope and *y*-intercept of this line are beyond the scope of this course. Teachers may encourage interested students to research these topics.

The least squares regression line is number 4 (or number 8) on the STAT , CALC menu. Students can calculate the least squares line using the same procedures described above for finding the median-median line.
The least squares regression line for this data, with slope and *y*-intercept rounded to four decimal places, is $y = .0375x + 8.8071$. The correlation coefficient, rounded to four decimal places is $r = .6703$. Students should draw this line on their plots using a third color.

7. Consider the three lines of fit that you have drawn on your scatterplot.
   a. In the case of each line, what does the slope represent? What does the *y*-intercept represent?
   The slope of each line is the rate of change that tells us how many additional votes we can expect for Buchanan given one additional vote for Perot. Using our "eyeball" line, we can say that for every additional vote for Perot, we can expect approximately .023 votes for Buchanan. However, given the decimal value of this slope, it may make more sense to say that for every 1000 votes for Perot, we can expect 23 votes for Buchanan. The *y*-intercept has no contextual value in this case. Since Buchanan's votes are always less than Perot's, it makes no sense to say that Buchanan receives 86 votes when Perot receives no votes.

   b. How are the lines alike? How are they different? Which line, in your opinion, appears to best fit the data? Explain your thinking.
   Students may note that all three lines are alike in that they show a positive association between the number of votes for Perot and the number of votes for Buchanan. Slopes and y- intercepts are similar. However, it is important to note that the slope of the least squares regression line is larger than the slope of the median-median line and, in our case, larger than the slope of the "eyeball" line. In the case of the median line, we expect 258 votes for Buchanan for each additional 10,000 votes for Perot. Using the least squares line, we expect 375 votes for Buchanan for every additional 10,000 votes for Perot. These slopes illustrate the fact that the least squares regression line is more affected by the regression outlier than the median-median line. The least squares regression line is "pulled" toward the outlier.

Students are asked to determine which line they feel best approximates more of the data. They may base their opinions on a simple visual inspection. They may also consider that a correlation coefficient of .67 tells us that the LSRL is not a great predictor of the votes for Buchanan, given votes for Perot.

c. Use each of your lines to predict the vote for Buchanan in the counties named in the table below. Use the table to organize your information.

	Predicted Vote for Buchanan			
	Liberty	Manatee	Hillsborough	Palm Beach
Eyeballing	95	324	665	793
Median-Median Line	47	305	687	831
Least Squares Regression Line	23	397	951	1160
Actual Vote for Buchanan	39	272	847	3407

This item should help students examine more closely the predictions provided by each of the regression lines. The calculator can be particularly helpful in finding these values quickly. If students have entered all three regression equations in the $Y =$ menu, they may use the VARS menu to enter *Y variable* (*x-value*) to calculate y. For example, to calculate the Buchanan vote for Liberty County using the median-median line, if the median were entered in Y_2, enter Y_2 (376) and press *ENTER*.

This is a very small set of data values and the set includes an influential outlier. There is no clear pattern here. Using the eyeball line $y = .023x + 86$ and comparing predictions from all three lines, the median-median line gives the closet prediction for Liberty and Manatee, where the x- values are relatively small and the second closest for Hillsborough. The LSRL gives the second best prediction for Liberty, the worst for Manatee, and the best for the last two counties where the x-values are fairly large.

8. The **residual** for any point in a data set is the **prediction error** that occurs when a regression line is used to predict the y value for the given value of x. If the data point lies above a regression line, its **residual** is defined to be the vertical distance between the line and the point. If the data point lies below the regression line, its residual is the opposite of the vertical distance from the point to the line. (Some refer to this as the *signed* distance from the line since the residual is positive when the data point is above the line and negative when the data point is below the line).

a. On your scatterplot, label each of the points representing the four counties in the table above with their ordered pairs.

b. Look again at the **least squares regression line** that you found in *Item 6c*. Mark the points on you least squares regression line that correspond to the counties in the table. For each point, draw a line segment that represents the vertical distance from the regression line to the corresponding data value in the scatterplot.

c. Find the residuals related to the least squares regression line for the points in the table.

Residuals			
Liberty	Manatee	Hillsborough	Palm Beach
16	- 125	- 104	2247

d. Explain what each of the residuals in *part c* tells you?
A residual of 16 for the point representing the vote in Liberty County tells us that there were actually 16 more votes for Buchanan than predicted by the LSRL. In other words, the prediction was under by 16 votes.

A residual of - 125 for the point representing the vote in Manatee County tells us that there were 125 less votes for Buchanan than predicted by the LSRL. The prediction was over by 125 votes.

A residual of - 104 for the point representing the vote in Hillsborough County tells us that there were actually 104 less votes for Buchanan than predicted by the LSRL. The prediction was over by 104 votes.

A residual of 2247 for the point representing the vote in Palm Beach County tells us that there were actually 2247 more votes for Buchanan than predicted by the LSRL. The prediction was under by 2247 votes.

e. Explain how would you compute the residual for any point in this data set?
At this point, we are asking students to be specific. How would they calculate the residual using the LSRL for any point in this particular data set? In the next item, we will generalize this idea to any data set.
The residual for any point in the data set, using the LSRL, can be computed by subtracting the predicted vote for Buchanan for a given county from the actual vote for Buchanan in that county.

The least squares regression line is often represented as $\hat{y} = ax + b$, where "*y*-hat" is the predicted value of *y* for a given *x*. As a result of the formula used to calculate the line, the following properties hold for the least squares line of any data set:

- The sum of the residuals is 0.
- The regression line includes the point $(\bar{x}, \bar{y})$.

- The regression line minimizes the sum of the squares of the residuals.
- There is a *unique* least squares regression line for a given set of data.

9. Consider that a data set has n distinct data points.
   a. Write an ordered pair for the i^{th} data point.
   $$(x_i, y_i)$$
   b. Using the symbol for "*y*-hat", write an ordered pair for the point on the least squares regression line corresponding to the i^{th} data point
   $$(x_i, \hat{y}_i)$$
   c. Using the ordered pairs from *parts a* and *b*, write an algebraic expression for the residual of the i^{th} data point.
   $$y_i - \hat{y}_i$$
   d. In Item 6, the least squares regression line is described as the line that minimizes the sum of the squares of the vertical distances between the line and the points in the data set. In other words, the least squares regression line minimizes the sum of the squares of the residuals. Write an algebraic expression for the sum of the squares of the residuals of a set of data with n distinct data points.
   $$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

10. Use your calculator to verify that the sum of squares of the residuals for the median-median line is greater than the sum of squares of the residuals for the regression line determined in *Item 5*. Which line do you think appears to be the better fit for this data? Why do you think this might be true?
    The calculator should be used to find the sum of squares of the residuals for both the median-median line and the LSRL. This may be done in the following manner.
    If Diagnostics is turned on, the residuals for a given set of data and the *last regression line (or curve) entered* can be found by pressing *2nd, STAT, 7, ENTER*. The residuals may then be stored in a list by pressing STO, List #, and ENTER. To obtain the sum of squares of the residuals, press *STAT*, *CALC*, and *1. 1-Var Stats* will appear on the screen. Press the appropriate list number and ENTER. The sum of the squares of the residuals will be noted as $\sum x^2$. The sum of squares of the residuals for the median-median line for our set of data is 6,947,568.57. The sum of squares of the residuals for the LSRL for our data is 6,190,426.69.
    The median-median line should appear to be a closer fit to more of the data, particularly those points with smaller *x*-values. This is because the median-median line is calculated using medians while the LSRL is calculated using means. Students should know from their work in this and previous courses that medians are more resistant to outliers than means.

11. A **regression** outlier is defined as a point that is not necessarily an outlier in x in relation to the other x-values or in y in relation to the other y-values but rather a point that is well removed from the trend of the other data. Are there regression outliers in our data? If so, what are the outliers?
    The data point for Palm Beach County is an outlier. While its x-value is not unusual, its y-value is high and the point is certainly well removed from the trend of the other data.

12. An influential outlier is a regression outlier with an x-value that is relatively high or low compared to the rest of the data. Influential outliers have a large effect on regression analysis. Are there influential outliers in your data? If so, recalculate the regression line without the influential outlier and then calculate the sum of squares of the residuals for your new regression equation. Compare this value to the sums of the residuals in problem 10. What does this comparison tell you?
    The data point for Palm Beach County (30739, 3407) is an influential outlier. The x-value of this point, while not an outlier compared to other x-values, is relatively large. Regression outliers with relatively high or low values for x, tend to pull the regression line in their direction more than an outlier that occurs near the "center" of a set of data.

    The regression line, calculated without the influential outlier (30739, 3407), has equation $y = .0233x + 64.4905$ with slope and y- intercept rounded to four decimal places. The correlation coefficient for this line is $r = .925$. The sum of squares of the residuals is 292,568 compared to approximately 6,190,427 for the previous LSRL.

    A comparison of both the correlation coefficients and the sums of squares of the residuals tell us that, without the influential outlier, there is a much stronger linear association between the number of votes cast in the Florida counties for Perot in 1996 and the number of votes cast for Gore in 2000. A correlation of .925, represents a strong positive linear relationship.

13. Use your new regression line to predict the votes for Buchanan in Liberty, Manatee, Hillsborough, and Palm Beach counties.

	Predicted Vote for Buchanan			
	Liberty	Manatee	Hillsborough	Palm Beach
Least Squares Regression Line calculated without the ordered pair (30,739, 3407)	73	306	651	781
Actual Vote for Buchanan	39	272	847	3407

14. As a statistician, what conclusions might you draw from the analyses you have completed in this task? Write a short paragraph explaining your thinking.
    Students might include some of the points listed below in their discussions.
    - Upon plotting the data points for the 25 counties represented, it is immediately apparent that the point representing Palm Beach County is well removed from the trend of the other data.
    - There appears to be a linear association between the number of votes cast for Perot in 1996 and the number of votes cast for Buchanan in 2000.
    - Examining this apparent association, we first obtained three different lines of fit (eyeball, median-median, and the least squares regression line).
      y = (answers will vary) "eyeball line"
      $y = .0258x + 37.6273$ median-median line
      $y = .0375x + 8.8071$ least squares regression line
      All three lines had similar slopes and y-intercepts.
      The correlation coefficient for the least squares regression line, calculated with the Palm Beach County vote, is approximately .67. This value of r indicates a linear correlation but not necessarily a strong correlation.
    - Calculating a least squares regression line *without* the Palm Beach County vote (without the influential outlier), we obtained the regression line $y = .0233x + 64.4905$.
      - The slope of this regression line is much closer to the slope of the median-median line calculated using *all* of the data. The median-median line predicts that for every additional 10,000 votes for Perot, Buchanan will receive 258 votes. The new regression line predicts that for every 10, 000 additional votes for Perot, Buchanan can expect to receive 233 votes.
      - The correlation coefficient for this line is $r = .925$.
      - The sum of squares of the residuals is 292,568 compared to approximately 6,190,427 for the previous LSRL.
      - A comparison of both the correlation coefficients and the sums of squares of the residuals tells us that, based on the least squares regression lines, there is a much stronger linear association between the number of votes cast in the Florida counties for Perot in 1996 and the number of votes cast for Gore in 2000. A correlation of .925, represents a strong positive linear relationship.
    - The information gained in this analysis indicates association between the number of votes cast for Perot and the number of votes cast for Buchanan. It makes sense to use either the median-median line or the least squares regression line, calculated *without* the Palm Beach vote, to make a reasonable **prediction** for what the vote in Palm Beach County might have been based on the strength of the association examined here. We can agree that the vote appears to be unusual. However, we cannot assume, based on this association, that a confusing ballot *caused* this unusual vote. Nor can we assume other causes without further investigation.

CALCULATOR OPERATIONS

A. TO PUT DATA IN A LIST

1. Turn calculator on.
2. Press the STAT button.
3. Press ENTER

If you want to erase a column of numbers:

- Use the up arrow to highlight List 1
- Press the CLEAR button
- Press ENTER

4. Start entering the data. Press ENTER after each number.
5. To get out of the list screen, press QUIT (2^{nd}, Mode).

B. TO GRAPH A SET OF DATA IN YOUR LISTS

1. Go to STAT PLOT (2^{nd}, y =).
2. If you want to turn off all the graphs:
   - Arrow down to 4, Plots Off.
   - Press ENTER, Press ENTER again. This turns off all plots.
3. To turn Plot 1 on, go to STAT PLOT. (2^{nd}, y =).
4. Press ENTER.
5. Press ENTER when the flashing curser is over the ON button. This turns Plot 1 on.
6. Arrow down to Type. Move the right or left arrow so that the curser is over the type of graph you want.
7. Press ENTER.
8. Use the down arrow to enter the list where the data is that you want to graph.
9. Type in the list (Example: L_1. Remember, to get L_1, you hit 2^{nd} and then 1.).
10. Press ENTER.
11. Press GRAPH.
12. If the graph does not fit the graphing window or you do not see the graph, you will have to change the window dimensions.
    - Go to WINDOW to resize the graph.
    - Type in the X-Min, X-Max, S scale, etc. that fit your data
    - Remember, the negative sign like in the number, – 4, must be entered with the key (–) that is located next to the ENTER button.

C. TO CALCULATE STANDARD DEVIATIONS, MEANS, OR OTHER STATISTICS

1. Enter your data in a list.
2. Press the STAT key.
3. Use the right arrow to highlight CALC. Press ENTER
4. The screen should say 1-VAR STATS.
5. Type in your list (Example: L_1. Remember, to get L_1, you hit 2^{nd} and then 1.).
6. Press ENTER.
7. Your calculations should be there.
8. Use the down arrows to get more statistics.

Calculator Operations

9. To get to an empty screen, called the home screen, press the QUIT key (2^{nd}, Mode).

D. TO WRITE AND GRAPH THE EQUATION OF A LINE

1. Since a line represents bivariate data, you must enter at least two sets of data into the lists in your calculator.
2. Enter your data into two lists. (See Section A for instructions how to do this.)
3. To write the equation of the Least Squares Regression Line (LSRL)
   - Hit the STAT button
   - Arrow over to CALC
   - Scroll down to option 8
   - Hit the ENTER button
   - Type in L_1, L_2 (2^{nd} 1, 2^{nd} 2), VARS (next to the CLEAR button), arrow to Y-VARS, FUNCTION, Y1, ENTER.
   - If this is done correctly, you should have the information that gives you the equation of the LSRL and this equation is also pasted into the calculator under Y = .
   - Press GRAPH, and the data points and the line should be graphed.

HELPFUL HINTS:

1. ZOOM 9, automatically scales the data to fit the viewing window.
2. If you want to see r and r^2, you can turn this function on by using the CATALOG button (above the number zero). Go to CATALOG, arrow down to DiagnosticOn, and then hit ENTER and then ENTER again. The screen should say Done.

E. TO CALCULATE THE RESIDUALS FOR THE DATA OF A LSRL

1. The calculator will automatically calculate the residuals (residuals = $y_i - \hat{y}_i$) after the equation for the line is written.
2. Go to LIST (this is above the STAT button)
3. Scroll down to RESID and then press ENTER and ENTER again. This gives you a list of the residuals, point by point.
4. The residuals can be put into a list by hitting the store button, STO, and then enter the list you want to store the residuals in. An example of this would be STO L_3.
5. The residuals can also be put into a list by going directly to lists, STAT, ENTER, put the cursor over an empty list, hit the LIST button and scroll down to RESID and then press ENTER and ENTER again.
6. A residual graph can be plotted as a graph. Go to STAT PLOT, Plot 1, and put the list you want for the X values and the RESID for the Y values. ZOOM 9, will automatically size the window for the residual graph.

MORE TI-83^{+}/84 CALCULATOR COMMANDS

LIST menu (access to all list names, various list commands, and math functions for lists)

1) **LIST ⇒ NAMES ⇒** various
   a) Use arrow keys and ENTER to select the desired list. The names will be pasted to wherever you are in the calculator. NOTE: to access L1 to L6, simply use the 2ND key and 1..6 keys)

2) **LIST ⇒ OPS ⇒5:seq**
   a) format **seq** (*expression, variable, begin, end, increment*)
   b) This is a useful command to lead a list with numbers. Here are some examples:
      i) **Seq** (X,X,1,10,1) **STO->** L1 loads list L1 with 1,2,3,4,5,6,7,8,9,10
      ii) **Seq** (X,X,2,20,2) **STO->** L2 loads list L2 with 2,4,5,8,10,…,18,20
      iii) **Seq** (X^2,X,1,10,1) **STO->** L3 loads L3 with 1,4,9,16,25,36, …, 81,100

DISTR menu:

1) **DISTR ⇒ DISTR ⇒ 2:normalcdf** The normal cumulative density function
   a) (format: **normalcdf**(*lower-bound, upper-bound, mean, std deviation*))
   b) This function calculates the area under a normal curve between the given bounds. This allows you to skip the process of calculating a z-score and using Table A from the book.
   c) For example: **normalcdf**(0,36,35,2) finds the area under the curve to the left of x=36 in normally distributed data with a N(35,2) distribution.
      i) NOTE: although 0 is theoretically under that curve, notice it is far outside of 3 standard deviations and suffices to keep the calculator from exploding. Make sure you choose an upper bound or lower bound (depending on if you need to go left or right of your x) that is more than 3 standard deviations away from the mean.
   d) If you do not enter a mean and standard deviation, the calculator uses mean = 0 and std. deviation = 1 as a default (notice that is the standardized normal curve!)
   e) **3:invNorm** format: **invNorm** (*cumulative area, mean, st.dev*)
      This allows you to find specific values for *x* using a normal distribution, given a specified cumulative area. Always enter the cumulative area as a proportion, not a percentage.

2) **DISTR ⇒ DRAW ⇒ 1:ShadeNorm**
   a) (format: **ShadeNorm**(*lower-bound, upper-bound, mean, std deviation*))
   b) You MUST adjust the WINDOW before graphing so you can view the distribution
   c) This function draws the normal curve based on *mean* and *std deviation* given and shades between *upper-bound* and *lower-bound*.
   d) If you do not enter a mean and standard deviation, the calculator uses mean = 0 and std. deviation = 1 as a default

3) **DISTR ⇒ DISTR ⇒ A:binompdf** (The binomial probability density function)
   1) Format: **binompdf** (*num-trials, prob. of success, x*)

2) This function will display the probability of x ONLY in the B(n, p) distribution.
a) For example, **binompdf**(2, .5, 1) will display P(x=1) for the binomial distributon with *n*=2 trials and *p*= probability of success = 0.5.
3) If you do not enter an "x" number, it will display all individual probabilities of the B(n,p) distribution.
   a) If you use this option, use STO-> to store the distribution in a list for easy viewing
   b) For example, **binompdf**(2,.5) STO-> L1 will store .25, .5, .25 in L1.
   c) REMEMBER the distribution starts with ZERO and not with one. L1(1) will say .25, which is P(x=0).

4) **DISTR ⇒ DISTR ⇒ A:binomcdf** (The binomial cumulative density function)
   1) Format: **binomcdf** (*num-trials, prob. of success, up to x*)
   2) This function will display the cumulative probabilities up to x for the B(n,p) distribution.
   3) For example, **binomcdf**(3, .5, 2) will display P(x=0) + P(x=1) + P(x=2) for the B(3, .5) distribution (.875)
   4) If you do not enter an "up to x" number, it will display all cumulative probabilities (x=0, (x=0 + x=1), (x=0 + x-=1 + x=2)) for our previous example.

5) **MATH menu:**
   MATH ⇒ PRB ⇒ 5:randInt (The random integer function)
   1) Format: randInt(lower-bound, upper-bound, quantity)
   2) This function produces "quantity" number of random integers between lower-bound and upper-bound.
   3) For example, randInt(1,10,5) will produce 5 random numbers between 1 and 10 inclusively.
   4) You can store your random integers by using the format: randInt(lower-bound, upper-bound, quantity), push STO->, then input a list (such as L1).

   Other functions to take note of:

   1:rand ⇒ produces a random, real number between 0 1and 1.
   6:randNorm(*mean, std. Deviaiton, quantity*) ⇒ produces "quantity" random numbers from the Normal Distribution described by *mean* and *std. Deviation.*
   7:randBin(*num-trials, prob. Of success, x*) ⇒ produces "x" random numbers from the Binomial Distribution described by "n" (*num-trials*) and "p" (*prob. Of success*).

6) **STAT menu:**
   TESTS **[Choose appropriate option for either forming a confidence interval or conducting test of significance.]**

ATLANTA PUBLIC SCHOOLS
Mathematics & Science Initiative
Making A Difference

Atlanta Public Schools

Teacher's Curriculum Supplement

Mathematics II: Unit 2

Task 2: Straight or Curved?

Mathematics II
Task 2: *Straight or Curved?* **Day 1/1**
(GaDOE TE Problems 1 - 3)

Standard(s): MM2D2. Students will determine an algebraic model to quantify the association between two quantitative variables.

a. Gather and plot data that can be modeled with linear and quadratic functions.
c. Understand and apply the process of linear and quadratic regression for curve fitting using appropriate technology.

New vocabulary: quadratic regression

Mathematical concepts/skills:

- determining whether there is an association between two variables
- examining the strength of the linear relationship between two variables as measured by the Pearson Correlation Coefficient
- using visual inspection and residuals to determine which regression curve, linear or quadratic, is a better approximation of the relationship between two variables represented by a set of data points
- using regression equations to make predictions
- using appropriate technology to conduct statistical analyses

Prior knowledge:

- plotting sets of data
- using graphs, specifically scatterplots, to examine sets of data
- evaluating quadratic functions for given values of the variable

Essential question(s): How can I determine which regression curve best approximates the relationship between two variables?

Suggested materials: graph paper, colored pencils, graphing calculators, rulers, Smartview or projection devices for technology (if possible)

Warm-up: Ask students to examine the chart provided at the beginning of the task and, working alone, complete *Problems 1* and *2*.

Opening: Discuss *Problems 1* and *2*, making sure that students understand the relationship being investigated. Scales used for the scatterplot should be large enough so that students can plot points fairly accurately. (See teacher notes beginning on page 45.)

Worktime: This task gives students an opportunity to revisit many of the concepts learned in the first task of this unit, provides additional practice in using the statistical operations of the TI graphing calculator, and introduces quadratic regression. Students should work in pairs or groups to complete all parts of the task.

It is important for students to draw the scatterplot of the data on graph paper and copy the regression line onto the plot as directed in *Problem 4*. After students have had ample time to complete *Problems 3* and *4*, have a whole class discussion of these items.

Problem 5 directs students to draw a scatterplot of the residuals for the data and the regression line. Many students will not realize that the *x*-coordinates of the points in this plot will be the *x*-values of the original data points. Ask guiding questions in order to remind students that a residual is the vertical distance between the data point and the regression line. The *x*-values of these points remain unchanged.

Have a whole class discussion of *Problems 5* and *6* before beginning *Problem 7*. Technology should be used to determine the sum of the squares of the residuals in *Problem 6*. This number will not be useful until students find the sum of squares of the residuals for the quadratic regression equation in *Problem 10*. At that point, students can compare the two numbers as a measure of fit.

Closing: Discuss *Problems 7 – 10,* thoroughly.

Homework:

Differentiated support/enrichment:

Check for Understanding:

Resources/materials for Math Support: Students may need extra practice in the following:

- understanding and plotting residuals
- using the statistical operations of the graphing calculator

Mathematics II
Task 2: *Straight or Curved?*
Student Task

World Travel	Mexico City	Cairo	London	Tokyo	Calcutta	Moscow	Rome
Air distance from New York City in miles	2094	5602	3458	6740	7918	4665	4281
Roundtrip fare from New York in dollars	250	750	375	1200	1500	624	520

1. Using the data provided above, we want to determine whether there is an association between air distance traveled from New York to a given destination and the roundtrip airfare charged for the flight.

    a. What is the appropriate explanatory variable for this situation? Be specific.

    b. What is the appropriate response variable?

2. Draw a scatterplot of the data on graph paper. What scales did you use?

3. Use technology to find a least squares regression line for the data.

    a. What does the slope of this line tell us in the context of this situation?

    b. What does the *y*-intercept tell us in this context?

    c. Use your line to predict the roundtrip airfare for a flight from New York City to Paris. The air distance between New York and Paris is 3636 miles.

    d. How strong is the linear association between the variables in this situation? Explain your thinking.

4. Copy the regression line onto your plot and describe how the data falls about the line.

5. We can more closely examine how data points fall about a regression curve (including a line) by drawing a scatterplot of the residuals for the data and the regression model.

    a. What should you use as the *x*-values for a scatterplot of the residuals of a set of data points?

    b. What are the *y*-values?

c. What are reasonable scales for this plot?

d. What is the residual for the point involving a flight from New York to London? What does this residual tell you? Be specific.

e. How does this scatterplot of the residuals compare to the graph completed in *Problem 4* (the scatterplot of the data together with the graph of the regression line)?

f. Describe any patterns that you see in the plot of the residuals. What does the plot tell you about how well the regression line fits the data?

6. Determine the sum of squares of the residuals for the regression line.

7. Use technology to find a quadratic regression equation for the data.

   a. Draw the quadratic function on your original plot of the data.

   b. Use your quadratic equation to predict the roundtrip airfare for a flight from New York City to Paris. The air distance between New York and Paris is 3636 miles.

8. Which model, linear or quadratic, do you think best fits the data? Explain your thinking.

9. Make a scatterplot of the residuals for your quadratic regression equation.

   a. What is the residual for the point involving a flight from New York to London? What does this residual tell you? Be specific.

   b. Describe any patterns that you see. What does the plot tell you about how well the regression equation fits the data?

10. Determine the sum of squares of the residuals for the quadratic regression equation. How does this value compare with the value you obtained in *Problem 6?* What do these numbers tell you?

Teacher Notes

Task 2: *Straight or Curved?*

This task gives students an opportunity to revisit many of the concepts learned in the first task of this unit, provides additional practice in using the statistical operations of the TI graphing calculator, and introduces quadratic regression.

World Travel	Mexico City	Cairo	London	Tokyo	Calcutta	Moscow	Rome
Air distance from New York City (miles)	2094	5602	3458	6740	7918	4665	4281
Roundtrip airfare from New York City (dollars)	250	750	375	1200	1500	624	520

1. Using the data provided above, we want to determine whether there is an association between air distance traveled from New York to a given destination and the roundtrip airfare charged for the flight.
   a. What is the appropriate explanatory variable for this situation? Be specific.
      Air distance (in miles) from New York City to the destination city is the appropriate explanatory variable.

   b. What is the appropriate response variable in this situation?
      Roundtrip airfare (in dollars) from New York City is the appropriate response variable.

2. Draw a scatterplot of the data on graph paper. What scales did you use?
   Answers will vary. It is reasonable to use a scale of 1000 miles for the explanatory variable and $100 for the response variable. Other scales are certainly acceptable. The scales should be large enough for students to plot the data fairly accurately.

3. Use technology to find a least squares regression line for the data.
   This problem gives students opportunities to revisit calculator operations including entering data into lists, graphing plots, determining regression curves, and automatically entering regression equations into the *Y*= menu. The least squares regression line for this set of data (with slope and *y*-intercept rounded to five decimal places) is $y = .22311x - 362.27662$.
   a. What does the slope of this line tell us in the context of this situation?
      The slope of the regression line predicts that for every additional 1 mile flown, the roundtrip airfare will increase by approximately $.22311. Students should be encouraged to make this statement several different ways. For example, for every 1000 additional miles, the airfare will increase by approximately $223.11.
   b. What does the *y*-intercept tell us in this context?
      The *y*-intercept has no meaning in this context. It makes no sense to predict that when a person does not fly (no miles traveled), the airfare is approximately - $362.28.

c. Use your line to predict the roundtrip airfare for a flight from New York City to Paris. The air distance between New York and Paris is 3636 miles.
Using the regression line, the predicted cost of the flight would be $448.96.

d. How strong is the linear association between the variables in this situation? Explain your thinking.
The linear association is strong based on a correlation coefficient of $r = .9743437108$.

4. Copy the regression line onto your plot and describe how the data falls about the line.
Despite the fact that the linear association between the variables in this situation is strong, the manner in which the data points fall about the regression line, indicates that there may be an even better fit to this data. Notice that three of the data points fall above the regression line. The *x*-values for these three points are relatively small or relatively large. In other words, they are on the "ends" of the plot. All of the data points between the first point and the last two (those in the "center" of the data) fall below the regression line. When a line is a good fit to data, we expect the data points to be distributed above and below the line as we move from left to right on the plot. When data points are distributed, as they are in this plot, with end points above the line and points in the center of the data below the line (or vice versa), there is a possibility that other regression curves, particularly quadratic models, may be a better fit to the data.

5. We can more closely examine how data points fall about a regression curve (including a line) by drawing a scatterplot of the residuals for the data and the regression model.
   a. What should you use as the *x*-values for a scatterplot of the residuals of a set of data points?
   In order to plot the residuals, we use the values of the explanatory variable as our *x*-values.
   b. What are the *y*-values?
   The residuals are the *y*-values for this plot.
   c. What are reasonable scales for this plot?
   Remember that to find residuals using the TI-83/84 families, students need to have Diagnostics turned On. Residuals should be found and stored in a list so that the sum of squares of the residuals can be found in the next problem. Although students can plot this graph on their calculators, it is important that they also draw this first plot of residuals by hand to reinforce understanding.
   The largest residual for this set of data is 145.08. The smallest is – 137.60. A scale of 1000 miles for the *x*-values is still appropriate. A scale of $10 is reasonable for the *y*-values. Other scales are certainly acceptable. If students use the same scale for the *y*-values that they used when plotting the original data values, they can more readily see how their graph of the original scatterplot together with the regression line compares to the plot of the residuals. (See *part d* below.)
   d. What is the residual for the point involving a flight from New York to London? What does this residual tell you? Be specific.
   Using the regression line to predict the roundtrip airfare from New York to London, we obtain a fare of $409.25 (rounded to the nearest cent). The actual fare is $375. This means that, using the regression line, we would predict a fare that is $34.25 more than the actual fare.

e. How does this scatterplot of the residuals compare to the graph completed in *Problem 4* (the scatterplot of the data together with the graph of the regression line)?
If we were to transpose the regression line so that it corresponds to the *x*- axis, the two graphs would look exactly the same. The plot of the residuals shows how far each data points lies above or below the regression line.

f. Describe any patterns that you see in the plot of the residuals. What does the plot tell you about how well the regression line fits the data?
Since the largest residual for this set of data is 145.08 and the smallest is – 137.60, the largest prediction error using the regression line would be about $145.08.

Students should notice the same pattern discussed in *Problem 4*. On the plot of the residuals, three points lie above the *x*-axis, indicating that the residuals are positive. The original data points corresponding to these three points lie above the regression line. Students should again notice that these points are the first and last two points moving from left to right. Residuals of points in the "center" of the data are negative, meaning the original data points corresponding to these residuals fall below the regression line. When a line is a good fit to data, ordered pairs representing the residuals should be distributed above and below the *x*-axis as we move from left to right on the plot. When ordered pairs representing residuals are distributed, as they are in this plot, with end points above the *x*-axis and points in the center of the data below the *x*-axis (or vice versa), there is a possibility that other regression curves, particularly quadratic models, may be a better fit to the data.

6. Determine the sum of squares of the residuals for the regression line.
The sum of squares of the residuals for the regression line, obtained using the TI-84 calculator, is 62,014.2622.

7. Use technology to find a quadratic regression equation for the data.
   a. Draw the quadratic function on your original plot of the data.
   The quadratic regression model for the data, obtained using the TI-84 calculator is $y = .000025x^2 - .032188x + 198.430221$. Coefficients have been rounded to six decimal places.
   b. Use your quadratic equation to predict the roundtrip airfare for a flight from New York City to Paris. The air distance between New York and Paris is 3636 miles.
   Using the quadratic regression model, the predicted cost of the flight would be $415.30.

8. Which model, linear or quadratic, do you think best fits the data? Explain your thinking.
By visual inspection, the parabola appears to be a closer fit to the data than the linear model. Five of the seven data points are very close to the curve. The parabola appears to be evenly centered between the two remaining points.

9. Make a scatterplot of the residuals for your quadratic regression equation.
    a. What is the residual for the point involving a flight from New York to London? What does this residual tell you? Be specific.

    Using the quadratic regression equation to predict the roundtrip airfare from New York to London, we obtain a fare of $389.14 (rounded to the nearest cent). The actual fare is $375. This means that, using the quadratic regression equation, we would predict a fare that is $14.14 more than the actual fare.

    b. Describe any patterns that you see. What does the plot tell you about how well the regression equation fits the data?

    Examining a plot of the residuals for the quadratic equation, students should see that points are distributed above and below the *x*- axis as the *x* values increase, indicating that the original data points are dispersed above and below the quadratic curve. The largest residual is 71.149 and the smallest is -60.74. This means that if the quadratic equation is used to predict airfare given the air distance traveled, the greatest prediction error would be about $71.15.

10. Determine the sum of squares of the residuals for the quadratic regression equation. How does this value compare with the value you obtained in *Problem 6?* What do these numbers tell you?

    The sum of the squares of the residuals for the quadratic regression equation is 10,443.321 compared to a sum of squares of 62,014.2622 for the linear equation. This measure of fit, given that there are no regression outliers in this set of data, would tell us that the quadratic model is a better fit to the data and therefore a better **predictor** of roundtrip airfare based on air distance traveled from New York City.

ATLANTA PUBLIC SCHOOLS
Mathematics & Science Initiative
Making A Difference

Atlanta Public Schools

Teacher's Curriculum Supplement

Mathematics II: Unit 6

Task 3: Orbital Debris

Mathematics II
Task 3: *Orbital Debris* **Day 1/1**
(GaDOE TE Problems 1 - 5)

Standard(s): MM2D2. Students will determine an algebraic model to quantify the association between two quantitative variables.

b. Gather and plot data that can be modeled with linear and quadratic functions.
d. Understand and apply the process of linear and quadratic regression for curve fitting using appropriate technology.

New vocabulary: interpolation, extrapolation

Mathematical concepts/skills:

- determining whether there is an association between two variables
- examining the strength of the linear relationship between two variables as measured by the Pearson Correlation Coefficient
- using visual inspection and residuals to determine which regression equation, linear or quadratic, gives a better approximation of the relationship between two variables represented by a set of data points
- using regression equations to make predictions
- interpolating using regression equations
- examining the feasibility of extrapolation using regression equations
- using appropriate technology to conduct statistical analyses

Prior knowledge:

- plotting sets of data
- using graphs, specifically scatterplots, to examine sets of data
- evaluating linear and quadratic functions for given values of the variable

Essential question(s): How can I determine which regression curve best approximates the relationship between two variables? When is it feasible to use regression equations to predict values?

Suggested materials: graphing calculators, Smartview or projection devices for technology (if possible)

Warm-up: Ask students to read the introduction of the task silently.

Opening: Discuss the introduction.

Worktime: This task reviews the concepts of linear and quadratic regression. The new content provided is related to interpolation and extrapolation and occurs in *Problems 4* and *5* Unless students continue to struggle with the concepts addressed in the first two tasks of this unit, it is not necessary for them draw the plots and graphs in this task by hand. Examining them on the graphing calculator or using statistical software should be sufficient.

Students should be able to complete *Problems 1 – 3* quickly. Monitor work carefully watching for those students still struggling with concepts addressed in the first two tasks and with using the graphing calculator.

When students have had ample time to complete *Problems 1 – 3*, have a whole class discussion of these items, stressing the definition and process of interpolation.

Discuss the definition of extrapolation provided in the student task and ask students to complete *Problems 4* and *5*.

Closing: Discuss *Problems 4* and *5,* thoroughly. (See GaDOE TE Teacher Notes.)

Homework:

Differentiated support/enrichment:

Check for Understanding:

Resources/materials for Math Support: Students may need extra practice in the following:

- understanding and plotting residuals
- using the statistical operations of the graphing calculator

Mathematics II
Task 3: ***Orbital Debris***
Student Task

Orbital Debris

Whenever a space shuttle is launched, mission-related debris are released as part of the mission. Satellites that have exhausted their missions remain in orbit, and the bodies of many of the rockets used to launch various spacecraft are still in space. Paint chips off spacecraft continue to float in orbit long after the spacecraft has returned to earth. All of these man-made items contribute to the debris in orbit about the Earth. The area where orbital debris is most concentrated is known as the LEO (low Earth orbit) and consists of space 200 and 2000 km above the Earth's surface. Telecommunications satellites are found in the geostationary orbit, above the LEO; according to the orbital debris specialists, there is little debris in the geostationary orbit.

Orbital debris can collide with current and future spacecraft and with each other. Although all spacecraft collide with small particles during a mission, a collision with a particle 10 cm in diameter (the measure across the widest part) or larger can cause catastrophic damage to a craft.

In response to the growing space pollution issue of orbital debris, the United States began close monitoring of orbital space debris, and, along with the United Nations, put into effect a plan to reduce the amount of orbital debris produced and left in space each year. In the United States, a plan for minimizing the creation of new orbital debris was proposed in 1997 and approved in 2001. One provision of this plan includes engineering space shuttle missions and satellite launches so that mission-related debris reenters the Earth's atmosphere and either burns up in descent through the atmosphere or falls safely to Earth in an uninhabited area.

All debris with a diameter of 10 cm or greater in low Earth orbit and with a diameter of 1 meter or greater in geostationary Earth orbit (GEO) are catalogued by the United States Space Surveillance Network. There are also hundreds of thousands of uncataloged smaller particles in Earth orbit. The work of monitoring and reporting orbital debris is carried out at the National Aeronautical and Space Administration (NASA) Orbital Debris Program Office at the Johnson Space Center in Houston, Texas. Scientists at the Johnson Space Center use sophisticated mathematical models and complex computer software to track orbiting debris and space collisions involving orbital debris. Tracking debris is a complex activity because the debris count changes so often. Throughout each year, new spacecraft are launched adding rocket bodies and other mission debris, and break-ups and collisions fragment large debris into many smaller pieces. While these processes add to the debris, some debris that has been in orbit for a long time falls back into the atmosphere to return to Earth or be burned up on reentry. In this task, we will model orbital debris data while we explore additional concepts about modeling data.

1. At the beginning of October 1997, there were 8545 man-made objects in Earth orbit. Over seventy percent of these were orbital debris: mission-related discards, rocket bodies, and fragmentation debris from break-ups and collisions. By late September 2006, there were 9800 man-made objects in Earth orbit, about seventy percent of which were debris.

   a. The table below lists the total number of orbital debris, as cataloged by the US Space Surveillance Network, for the years listed in the table[1].

Year	1997	1998	1999	2000	2001	2004	2005	2006
Man-made Orbital Debris (number of objects)	6166	6132	5980	6134	6092	6307	6480	6791

   The first man-made object to orbit the Earth was Sputnik 1, launched by the Union of Soviet Socialist Republics (USSR) on Oct. 4, 1957. The debris totals provided here are from dates near the end of the third quarter of each year. Thus, the dates for the data represent a whole number of years from the Sputnik 1 launch, and we can simplify our data by using the number of years since the launch as the first coordinate. Write points corresponding to the data in the table using the form (number of years since Sputnik 1 launch, number of man-made orbital debris).

   b. Make a scatterplot of the points from *part a*.

   c. Explain any trends in the data.

2. Use a calculator or other appropriate technology to find the correlation coefficient, *r*, for your data points from *Item 1*.

   a. Based on the value of the correlation coefficient, what are the strength (weak, moderate, or strong) and direction (positive or negative) of any linear relationship for this set of data?

   b. Using appropriate technology (graphing calculator, Excel, TI-Interactive, etc.), determine the least-squares regression line for the data.

   c. Graph the least squares regression line from *part b* together with a scatterplot of the data.

[1]Data taken from October issues of the Orbital Debris Quarterly Newsletter available online at http://orbitaldebris.jsc.nasa.gov/newsletter/newsletter.html . According to Nicholas L. Johnson, Chief Scientist for Orbital Debris, NASA Johnson Space Center, via email reply in November 2008, this data was current at the time of the newsletter printing but some of it was updated later because several months or even years may pass before all the data from a satellite fragmentation is detected and cataloged.

d. How well does the least squares regression line appear to fit the data?

e. Find the sum of the squares of the residuals between points on the regression line and the corresponding data points.

f. Interpret the slope of your least squares regression line from *part b*.

g. Interpret the *y*-intercept of your squares regression line from *part b*.

Many data sets studied by scientists and engineers, like the one studied in *Items 1* and *2*, can be viewed as functions since each input has a unique output. However, such real-life data are usually not modeled exactly by any function that we can build using basic functions or transformations of basic functions. However, mathematical models built from transformations of basic functions can provide useful information about the data. Thus, finding and comparing different models is a useful endeavor.

Linear functions are not the only functions we can use to model data. In the remainder of this task, we explore using quadratic functions to model data, called ***quadratic regression***. Modeling data with other polynomial functions and exponential functions is also possible. Although our study is limited to quadratic regression, we will see concepts that generalize for higher degree polynomials.

3. This item explores quadratic regression for the data points from *Item 1*. Quadratic regression finds the quadratic function that minimizes the sum of the squares of the vertical distances (residuals) between points on the quadratic graph and the data points.
   a. Use a calculator or other appropriate technology to find the quadratic regression equation for the data.

   b. Graph the quadratic regression curve from *part b* together with a scatterplot of the data.

   c. How well does the parabola appear to fit the data?

   d. Find the sum of the squares of the vertical distances (residuals) between points on the quadratic regression curve and the corresponding data points.

   e. Which function, the least squares linear regression or the quadratic regression, provides a better model for the data from *Item 1*? Explain your answer.

   f. The data in *Item 1* were obtained from the quarterly news letter published by the NASA Orbital Debris Program Office. The newsletter was not published for a period including the third quarters of 2002 and 2003 so, for these two years, the count of man-made debris in Earth orbit at the end of the third quarter was not available. Use your choice of the better model (from *part e*) to predict the amount of debris for these two years.

Finding the answers for *Item 3, part f* is an example of **interpolation**. Interpolation is used to make predictions within the domain of values of the independent variable. It is standard to find a regression curve for the purpose of finding one or more interpolated values.

Extrapolation is the use of a regression curve to make predictions outside the domain of values of the independent variable. Care must be taken to determine the feasibility of extrapolation; that is, sufficient evidence should exist that the trend described by the regression curve would continue outside the present domain.

4. Consider using your regression models for the data in *Item 1* to estimate man-made orbital debris during the period 1957 – 1966.

    a. Discuss whether it is feasible to use your least squares linear regression to extrapolate values for 1957 – 1966.

    b. Discuss whether it is feasible to use your quadratic regression function to extrapolate values for 1957–1966.

5. The amount of man-made debris can change in an instant when a satellite breaks apart or there is a collision among two pieces of debris. Approximately 1500 new debris were added in 2007 as a result of the breakup of the Fengyun-1C and Upper Atmospheric Research Satellites; these events made 2007 the worst ever year for producing new man-made debris.

    a. The actual count of man-made debris in Earth orbit at the end of the third quarter of 2007 was 9250. State the error in using your linear and quadratic models to predict the count of man-made debris in Earth orbit at the end of the third quarter of 2007.

    b. The actual count of man-made debris in Earth orbit at the end of the third quarter of 2008 was 9661. State the error in using your linear and quadratic models to predict the count of man-made debris in Earth orbit at the end of the third quarter of 2008.

    c. If you found a new linear or quadratic regression function using the data from *parts a and b* above as well as the data from *Item 1, part a,* would you be confident in using your model above to predict the amount of man-made debris in Earth orbit for 2010? Explain.

ATLANTA PUBLIC SCHOOLS
Mathematics & Science Initiative
Making A Difference

Atlanta Public Schools

Teacher's Curriculum Supplement

Mathematics II: Unit 6

Task 4: Causation versus Correlation

Mathematics II
Task 4: *Causation versus Correlation* **Day 1/1**
(GaDOE TE *Problems 1* and *2*)

Standard(s): MM2D2. Students will determine an algebraic model to quantify the association between two quantitative variables.

d. Investigate issues that arise when using data to explore the relationship between two variables, including confusion between correlation and causation.

New vocabulary: causation, controlled study, observational study, lurking variable

Mathematical concepts/skills:

- understanding the difference between correlation and causation
- realizing that confusion between correlation and causation is pervasive in popular media

Essential question(s): What questions should I ask to help determine the validity of conclusions drawn from data and reported in the media?

Suggested materials: technology for internet searches and introductory statistics textbooks

Warm-up: Students should have been assigned the reading that accompanies this task as a homework assignment prior to the lesson. Ask students to jot down 3 or 4 key concepts they learned from the reading and be prepared to discuss those ideas.

Opening: Allow students to discuss the key concepts presented in the article on correlation versus causation. The following ideas should be included in the discussion.

- Correlation and causation are different. Correlation indicates that there is a linear association between two variables. We can closely predict the value of the response variable (y-value) when we know the value of the explanatory variable (x-value). Causation means that one action or occurrence *causes* another.
- Correlation *never* proves causation.
- When two occurrences are correlated, it is often true that the two variables are both affected by one or more common causes. These common causes may be unstudied or unknown by researchers and are referred to as lurking variables.
- While there are many tools for establishing statistically significant correlation, it is very difficult to prove causality between two correlated events.
- Confusion between correlation and causation is a widespread statistical error, particularly pervasive in popular media.

Guiding questions might include:

- What is the difference between correlation and causation?
- Why is it easy to confuse correlation and causation? Give examples.
- What tools used to establish correlation have we studied so far?
- How do scientists prove causation?
- What are the differences between controlled studies and observational studies?

Worktime: Students should work in pairs or small groups to complete *Problems 1* and *2* of the task. Monitor student work carefully, watching for particularly interesting topics to be shared during the closing.

Accountability for this work will be important in keeping students "on track". Students should be required to produce quality work that can be turned in and assessed.

Closing: Allow as many groups of students to share as time allows.

Homework:

Differentiated support/enrichment:

Check for Understanding:

Resources/materials for Math Support:

Mathematics II
Task 4: *Causation versus Correlation*
Reading

Causation versus Correlation

STATS is a non-profit, non partisan research organization affiliated with George Mason University. The following is reproduced from their website (http://www.stats.org/faq_vs.htm.

One of the most common errors we find in the press is the confusion between *correlation* and *causation* in scientific and health-related studies. In theory, these are easy to distinguish — an action or occurrence can *cause* another (such as smoking causes lung cancer), or it can *correlate* with another (such as smoking is correlated with alcoholism). If one action causes another, then they are most certainly correlated. But just because two things occur together does not mean that one caused the other, even if it seems to make sense.

Unfortunately, our intuition can lead us astray when it comes to distinguishing between causality and correlation. For example, eating breakfast has long been correlated with success in school for elementary school children. It would be easy to conclude that eating breakfast *causes* students to be better learners. It turns out, however, that those who don't eat breakfast are also more likely to be absent or tardy — and it is absenteeism that is playing a significant role in their poor performance. When researchers retested the breakfast theory, they found that, independent of other factors, breakfast only helps undernourished children perform better.

Many studies are actually designed to test a correlation, but are suggestive of "reasons" for the correlation. People learn of a study showing that "girls who watch soap operas are more likely to have eating disorders" — a correlation between soap opera watching and eating disorders — but then they incorrectly conclude that watching soap operas *gives* girls eating disorders.

In general, it is extremely difficult to establish causality between two correlated events or observances. In contrast, there are many statistical tools to establish a statistically significant correlation.

There are several reasons why common sense conclusions about cause and effect might be wrong. Correlated occurrences may be due to a common cause. For example, the fact that red hair is correlated with blue eyes stems from a common genetic specification which codes for both. A correlation may also be observed when there is causality behind it — for example, it is well-established that cigarette smoking not only correlates with lung cancer, but actually causes it. But in order to establish cause, we would have to rule out the possibility that smokers are more likely to live in urban areas, where there is more pollution — or any other possible explanation for the observed correlation.

In many cases, it seems obvious that one action causes another. However, there are also many cases when it is not so clear (except perhaps to the already-convinced observer). In the case of soap-opera watching anorexics, we can neither exclude nor embrace the hypothesis that the television is a cause of the problem — additional research would be needed to make a

convincing argument for causality. Another hypothesis is that girls inclined to suffer poor body image are drawn to soap operas on television because it satisfies some need related to their poor body image. Yet another hypothesis is that neither causes the other, but rather there is a common trait — say, an overemphasis on appearance by the girls' parents — that causes both an interest in soap operas and an inclination to develop eating disorders. None of these hypotheses are tested in a study that simply asks who is watching soaps and who is developing eating disorders, and finding a correlation between the two.

How, then, does one ever establish causality? This is one of the most daunting challenges of public health professionals and pharmaceutical companies. The most effective way of doing this is through a *controlled study*. In a controlled study, two groups of people who are comparable in almost every way are given two different sets of experiences (such as one group watching soap operas and the other game shows), and the outcome is compared. If the two groups have substantially different outcomes, then the different experiences may have caused the different outcome.

There are obvious ethical limits of controlled studies – it would be problematic to take two comparable groups and make one smoke while denying cigarettes to the other in order to see if cigarette smoking really causes lung cancer. This is why epidemiological (or observational) studies are so important. These are studies in which large groups of people are followed over time, and their behavior and outcome is also observed. In these studies, it is extremely difficult (though sometimes still possible) to tease out cause and effect, versus a mere correlation.

Typically, one can only establish correlation unless the effects are extremely notable *and* there is no reasonable explanation that challenges causality. This is the case with cigarette smoking, for example. At the time that scientists, industry trade groups, activists and individuals were debating whether the observed correlation between heavy cigarette smoking and lung cancer was causal or not, many other hypotheses were considered (such as sleep deprivation or excessive drinking) and each one dismissed as insufficiently describing the data. It is now a widespread belief among scientists and health professionals that smoking does indeed *cause* lung cancer.

When the stakes are high, people are much more likely to jump to causal conclusions. This seems to be doubly true when it comes to public suspicion about chemicals and environmental pollution. There has been a lot of publicity over the purported relationship between autism and vaccinations, for example. As vaccination rates went up across the United States, so did autism. However, this correlation (which has led many to conclude that vaccination causes autism) has been widely dismissed by public health experts. The rise in autism rates is likely to do with increased awareness and diagnosis, or one of many other possible factors that have changed over the past 50 years.

In general, we should all be wary of our own bias; we like explanations. The media often concludes a causal relationship among correlated observances when causality was not even considered by the study itself. Without clear reasons to accept causality, we should only accept correlation. Two events occurring in close proximity does not imply that one caused the other, even if it seems to makes perfect sense.

Mathematics II
Task 4: *Causation versus Correlation*
Student Task

Do the following activities with the aid of research on the topic of correlation versus causation. An Internet search on this topic will produce a wealth of examples to consider. The introductory statistics textbooks, provided by your teacher, will also have information on the topic.

1. Find one example, or collect data to create an example of your own, of a strong correlation (either positive or negative) between variables where the idea that either one causes the other is clearly not reasonable.

   a) Describe the two variables that are correlated.

   b) Give the correlation coefficient or evidence to support the claim that there is a strong correlation.

   c) Write a sentence making the claim that one variable causes the second.

   d) Give an explanation of why the correlation may exist.

   e) Make a bibliography of all sources used.

2. Find an example of a topic of current public interest where there is a correlation between two quantities, call them A and B, that has been interpreted, at least in some media, as A causing B.

   a) State the claim that A causes B providing the source of the claim and explanations of terms as necessary.

   b) Discuss possible reasons that there may be a correlation without A actually causing B, such as a third variable C related to A and B.

   c) Make a bibliography of all sources used.