

Summer 2013

# Revising Common Core Georgia Performance Standards Statistics Lesson Plans to Better Align with Statistical Practice

Rachel Bonilla

Follow this and additional works at: <https://digitalcommons.georgiasouthern.edu/etd>



Part of the [Applied Statistics Commons](#), and the [Categorical Data Analysis Commons](#)

---

## Recommended Citation

Bonilla, Rachel, "Revising Common Core Georgia Performance Standards Statistics Lesson Plans to Better Align with Statistical Practice" (2013). *Electronic Theses and Dissertations*. 833.

<https://digitalcommons.georgiasouthern.edu/etd/833>

This thesis (open access) is brought to you for free and open access by the Graduate Studies, Jack N. Averitt College of at Digital Commons@Georgia Southern. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons@Georgia Southern. For more information, please contact [digitalcommons@georgiasouthern.edu](mailto:digitalcommons@georgiasouthern.edu).

REVISING COMMON CORE GEORGIA PERFORMANCE STANDARDS  
STATISTICS LESSON PLANS TO BETTER ALIGN WITH STATISTICAL  
PRACTICE

by

RACHEL NAVARRO BONILLA

(Under the Direction of Patricia B. Humphrey)

ABSTRACT

In this thesis, lesson plans provided by the Georgia Department of Education are revised to give students better exposure and practice working with real-life data. Three learning tasks and a performance task are presented covering a unit lesson on statistical regression. The development of Georgia statistics curriculum standards are reviewed and presented.

INDEX WORDS: Common Core Georgia Performance Standards, Regression, Statistics

REVISING COMMON CORE GEORGIA PERFORMANCE STANDARDS  
STATISTICS LESSON PLANS TO BETTER ALIGN WITH STATISTICAL  
PRACTICE

by

RACHEL NAVARRO BONILLA

B.S., University of Georgia, 2010

B.S., University of Georgia, 2011

B.S.Ed, University of Georgia 2011

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial  
Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

STATESBORO, GEORGIA

2013

© 2013

RACHEL NAVARRO BONILLA

All Rights Reserved

REVISING COMMON CORE GEORGIA PERFORMANCE STANDARDS  
STATISTICS LESSON PLANS TO BETTER ALIGN WITH STATISTICAL  
PRACTICE

by

RACHEL NAVARRO BONILLA

Major Professor: Patricia B. Humphrey  
Committee: Sharon Taylor  
Gregory Chamblee

Electronic Version Approved:  
July, 2013

## DEDICATION

I dedicate this thesis to my former math teacher and sister, Julie Arasato. Through her teachings and guidance, she has given me the strong mathematical foundation that has gotten me this far in my educational career as well as ignited my passion for teaching. She is my greatest role model, and I can only hope one day to be as great of a teacher as her inside and outside the classroom.

I also dedicate this thesis to my mom, who was the first person that introduced the area of statistics to me, and gave me the statistics gene. For all the sacrifices she has made to give me a bright future, while instilling in me good morals and values. I can only hope one day to be as great of a mom as she has been to me.

## ACKNOWLEDGMENTS

I wish to acknowledge my advisor Dr. Pat Humphrey for her continuous guidance, patience and understanding through my thesis process. I express my greatest gratitude to her for the countless hours she has dedicated to providing me the guidance necessary for the success of this thesis. Without her support, this thesis would not have been possible.

My special thanks go to my committee members, Dr. Sharon Taylor, and Dr. Greg Chamblee for the time they have given to me. It is my honor to have them in my committee, and I'm very grateful for the ideas and suggestions they have contributed to me for this thesis.

I would also like to thank my research professor Dr. Robert Mayes for giving me the great opportunity of being his research assistant the past two years. Thank you for all the invaluable learning experiences. Thank you to the Office of Research and Development for funding my assistantship.

I give my deepest gratitude to my family for their continuous support throughout my entire educational career. My success has been made possible through their constant encouragement and support.

To my academic brothers, Fidelis Mutiso, Benson Kimitei, and Satbir Malhi, I thank you for all the countless hours of learning and fun together. The past two years would not have been nearly as enjoyable as it was without you guys.

Lastly, thank you to all my professors, teachers and friends who have supported me and given me strength through the years and motivated me to reach for my dreams.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS .....	VI
LIST OF TABLES .....	IX
LIST OF FIGURES .....	X
CHAPTER	
1 INTRODUCTION .....	1
1.1 A Brief Historical Overview .....	1
1.2 Tracking Statistics Standards .....	3
1.3 Motivation for Revision of Lesson Plans .....	8
2 LEARNING TASK 1: SIMPLE LINEAR REGRESSION .....	11
2.1 Teaching Guide .....	12
2.2 Student Worksheet: Simple Linear Regression .....	26
2.3 Comparison Summary .....	30
3 LEARNING TASK 2: TV WATCHING HOURS AND TEST GRADES.....	31
3.1 Teaching Guide .....	32
3.2 Student Worksheet: TV Watching Hours and Test Grades .....	41
3.3 Comparison Summary .....	47
4 LEARNING TASK 3: U.S. POPULATION GROWTH.....	48
4.1 Teaching Guide .....	49
4.2 Student Worksheet: U.S. Population Growth .....	52
4.3 Summary .....	56
5 PERFORMANCE TASK: SAVINGS ACCOUNTS.....	57
5.1 Teaching Guide .....	58



5.2 Student Performance Task Worksheet .....	60
5.3 Comparison Summary .....	65
6 CONCLUSION .....	66
REFERENCES .....	67
APPENDICES	
A LEARNING TASK 1: SPAGHETTI REGRESSION .....	68
B LEARNING TASK 2: TV/TEST GRADES .....	78
C PERFORMANCE TASK: EQUAL SALARIES FOR EQUAL WORK? .....	88
D STUDENT WORKSHEET: LEARNING TASK 1 .....	99
E STUDENT WORKSHEET: LEARNING TASK 2 .....	103
F STUDENT WORKSHEET: LEARNING TASK 3 .....	107
G STUDENT WORKSHEET: PERFORMANCE TASK.....	109

## LIST OF TABLES

Table 1.1: Algebra 1 QCC vs Statistics and Probability GPS .....	4
Table 1.2: Interpreting Categorical and Quantitative Data .....	6
Table 1.3: Conditional Probability and the Rules of Probability .....	7
Table 1.4: Making Inferences and Justifying Conclusions .....	8
Table 2.1: Learning Task 1 Standards .....	12
Table 2.2: TV Time and Test Score .....	14
Table 2.3: Measuring Residuals .....	15
Table 2.4: TV Time and Test Score .....	26
Table 3.1: Learning Task 2 Standards .....	32
Table 3.2: Education and Income .....	34
Table 3.3: Education and Income Data .....	37
Table 3.4: Ms. Garth’s Class Data .....	42
Table 3.5: Correlation of Variables .....	43
Table 4.1: Learning Task 3 Standards .....	49
Table 4.2: Partial Population Data .....	50
Table 4.3: U.S. Population Data .....	52
Table 5.1: Performance Task Standards .....	58
Table 5.2: Savings Chart .....	60
Table 5.3: Savings Accounts .....	60

## LIST OF FIGURES

Figure 2.1: Best Fit Line Example .....	16
Figure 2.2: Overhead 1 .....	17
Figure 2.3: Overhead 2 .....	18
Figure 2.4: Measuring Residual Distances .....	19
Figure 2.5: Residual Definition.....	20
Figure 2.6: TV Hours vs Test Score with Best-Fit Line .....	21
Figure 2.7: Random Residual Plot .....	22
Figure 2.8: Non-Random Residual Plot.....	22
Figure 2.9: Calculator Diagnostic .....	23
Figure 2.10: Calculator Listing .....	24
Figure 2.11: Data Lists .....	24
Figure 2.12: Linear Regression .....	25
Figure 2.13: Stat Plots .....	25
Figure 2.14: Zoom Stat.....	25
Figure 2.15: Hours of TV vs Test Score .....	27
Figure 2.16: Linear Regression Output .....	28
Figure 2.17: Hours of TV vs Test Score Residual Plot.....	29
Figure 3.1: Correlation .....	35
Figure 3.2: Calculator Diagnostic.....	37
Figure 3.3: Calculator Listing.....	37
Figure 3.4: Data Lists.....	38
Figure 3.5: Stat Plots .....	38
Figure 3.6: Zoom Stat .....	39

Figure 3.7: Linear Regression .....	39
Figure 3.8: R-Squared Example .....	40
Figure 3.9: Correlation Calculator Outputs .....	44
Figure 4.1: Partial Population Data Scatter Plot .....	51
Figure 4.2: Population Scatter Plot .....	53
Figure 4.3: Scatter Plot Calculator Output.....	53
Figure 4.4: Linear Regression Calculator Output .....	53
Figure 4.5: Exponential Regression Calculator Output .....	54
Figure 4.6: Pre-War Regression Models and Residual Plot .....	55
Figure 4.7: Pre-War Scatter Plot with Model Fits .....	55
Figure 4.8: Post-War Regression Models and Residual Plot .....	55
Figure 4.9: Post War Scatter Plot with Model Fits .....	55
Figure 5.1: Savings Accounts Scatter Plot.....	61
Figure 5.2: John’s Linear Model .....	63
Figure 5.3: Lucy’s Linear Model .....	63
Figure 5.4: Mark’s Linear Model.....	63
Figure 5.5: Residual Plots.....	64

# CHAPTER 1

## INTRODUCTION

The importance of statistics education has been in the forefront of educational reform in recent years. It has been shown that citizens are constantly presented with data and numbers they need to understand in order to make informed decisions regarding their personal lives. As times have changed with technological advancement, today's society is dependent upon the collection and interpretation of data. Researchers have referred to this understanding as Quantitative Literacy. According to Jerry Moreno, assistant professor emeritus of statistics, "surely, a citizen should be able to read a newspaper intelligently, make decisions based on logic and quantitative information regarding political candidates, medicines and health, investments. Practically everyone in the workplace from farmers to lawyers, jurors to the accused, manufacturers to consumers' needs to be able to think quantitatively" (Moreno). Students need to be equipped with the tools necessary to have a general understanding of decisions and conclusions that are based on statistical evidence.

### 1.1 A Brief Historical Overview

Georgia mandated a state curriculum that specifies what students should know for each subject in each grade through the Quality Basic Education (QBE) Act of 1985. Georgia's state curriculum began with the creation of the Quality Core Curriculum (QCC) to provide a guideline of minimum standards which each school system must cover in the classroom to prepare their students for the state's standardized tests which were also required by this act.

In 2002, it was concluded by the group Phi Delta Kappa – a professional association for educators, - that the QCC lacked depth, could not be covered in a reasonable amount of time, and did not meet national standards. Criticism for QCC included shallow standards, which left teachers guessing what to teach and hoping that they had covered everything that will be on the standardized test. Teachers merely used the curriculum by mentioning them in the lesson plans, but nothing more (GA DOE). Because of this, the Georgia Performance Standards (GPS) were created by teachers, state and national experts, and consultants. Guidelines by National Council of Teachers of Mathematics (NCTM) and the American Association for the Advancement of Science (AAAS) were used as a guide; standards from high performing states such as Texas and North Carolina and countries such as Japan were also considered. GPS increased the depth of topics across content areas and provided instructors with suggested tasks. It defined expectations of skills students are to develop, acceptable assessment and instruction.

While the GPS was being developed, the No Child Left Behind Act (NCLB) was being imposed on states. This act required states to develop assessments in basic skills that are required to be given to students at three different grade levels in order to receive federal school funding. The aim of NCLB was to support standards-based education reform such as Georgia's development of the GPS, and increase accountability by

ensuring states standards allow equal opportunity for all students to receive a high school degree. With this in mind, the development of the GPS was made to align with the need to follow the NCLB Act (NCLB).

Achieve – a bipartisan, non-profit organization that helps states raise academic standards was founded in 1996. This organization is dedicated to supporting standards-based education reform across the nation. In 2005, Achieve; in partnership with the National Governor’s Association (NGA), sponsored the National Education Summit on High Schools. Forty five governors, along with corporate CEOs and education leaders from K-12 and higher education attended the summit. At the end, 13 states launched the American Diploma Project (ADP) Network with the goal of preparing students to be career and college ready. As of 2012, 35 states are part of the ADP network, educating 85% of the nation’s public school students. (Achieve) The motivation for creation of the Common Core State Standards (CCSS) arose from ADP. The states in the ADP network are making the “case that education and opportunity are critical to America’s ability to innovate, compete and grow in an increasingly sophisticated and technologically-driven world economy” (Achieve, 2012). With the CCSS, high standards are created to be consistent across states, ensuring all students are equally well prepared with the skills and knowledge necessary for success in college and their careers.

The nation’s governors and education commissioners through NGA and the Council of Chief State School Officers (CCSSO) led the development of the Common Core State Standards (CCSS). These standards are “internationally-benchmarked; college and career ready; rigorous, clear and focused, and grounded in research” (Achieve, 2012). The federal government was not involved in the development of the standards. CCSS was created through the collaboration of teachers, researchers and experts in curriculum design and development among the states. Beginning in 2010, each of the states independently chose to adopt the CCSS. On July 8, 2010 Georgia adopted the CCSS with a plan for full implementation during the 2012-13 school year. In 2009, 48 states, 2 territories and the District of Columbia signed an agreement committing to the CCSS Initiative (Achieve, 2013). Today, forty-five states have adopted the CCSS along with the District of Columbia and four territories. CCSS does not cover all subjects, but covers English language arts and mathematics as these are the subjects that are most frequently assessed and the basis of all other subjects (CCSSI).

Georgia’s formal adoption of the CCSS has resulted in the creation of the Common Core Georgia Performance Standards (CCGPS). This was created through a combination of the CCSS and the GPS. The purpose of CCGPS is to ensure that all Georgia students have equal access and opportunity to obtain the skills and knowledge necessary for success beyond high school. The focus of this thesis will be on the CCGPS Mathematics in grades 9-12.

The QBE act required Georgia to have state standardized tests. There are two types of standardized assessments, the Criterion Referenced Competency Test (CRCT) for grades 1 – 8, first implemented in 2002, and the Georgia High School Graduation Test (GHS GT) typically first taken in grade 11. Students who entered 9<sup>th</sup> grade between 1981 – 1991 had their GHS GT as the Basic Skills Test (BST). The GHS GT was based on the QCC curriculum between 1991 – 2008. This was also a transition period in which

QCC GHS GT was phased out and the GPS GHS GT was phased in, and became just GPS GHS GT between 2008 – 2011. Students who enter 9<sup>th</sup> grade after 2011 are no longer required to pass the GHS GT. (GA DOE) Since the creation of CCSS, the states that have adopted CCSS are to give a national assessment based on this curriculum. There are two assessment consortiums; Georgia is a member of the Partnership for Assessment of Readiness for College and Career (PARCC). Georgia will implement this national common assessment starting the 2014 - 2015 school year (Achieve, 2012).

## 1.2 Tracking Statistics Standards

With an understanding of the history of the mathematics curriculum, now, consider the grades 9-12 standards progression through the years with a focus on statistics standards present in the QCC, GPS and CCGPS. According to Lynn Arthur Steen, Professor Emeritus of Mathematics at St. Olaf College, “It is no accident that mathematics is the discipline that launched the standards movement in the United States. Mathematics is central to the global movement to democratize education: it undergirds the increased need for postsecondary education, technical demands of the high performance workplace, and conceptual skills required to live and work in the information age” (Steen). The addition of statistics in the mathematics standards began in the early 1980’s with the NSF-Funded Quantitative Literacy Project (QLP). This was a joint project of the American Statistical Association (ASA) and the National Council of Teachers of Mathematics (NCTM). The results of the project were used as a basis for the statistics strands in the NCTM standards and included curriculum materials (Scheaffer, 1990). NCTM included statistics in their Curriculum and Evaluation Standards in 1989, and in the Principles and Standards in 2000. In 2005, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report for Pre-K through 12 was adopted by the American Statistical Association and became the basis for statistics in the Common Core Standards (Franklin). This growth in development of statistics in the curriculum can be seen by the AP Statistics exam, first offered in 1997 with 7,500 exams given, growing to 171,097 exams given in 2013 (Humphrey).

The QCC math standards were created with respect to each mathematics course that could be offered by Georgia’s schools. The standards were separated into 20 courses. This included a wide variety of courses ranging from common courses such as algebra, geometry, and pre-calculus to more concentrated options such as analysis, and discrete mathematics. In particular, two courses focused on statistics were included – Statistics, and Concepts of Probability and Statistics. Within each of these courses, a list of topics to be covered is given along with the standards to be addressed with the topic. Many of the courses had about 40 standards, which is one of the reasons why QCC was criticized as being too broad and not able to be addressed within the course time allotment.

Thus, the GPS was formed to address such issues. GPS math standards for grades 9-12 were separated into four courses, Math I-IV. These courses are considered to be integrated due to the fact that each of the courses in the sequence consists of topics previously separated, such as algebra and geometry, blended together. All of the four

courses have content standards by topic, as well as process standards to be applied to each of the topics covered within the course. All four courses include Data Analysis and Probability as one of the topic strands to be covered in the course. Georgia was a visionary for having statistics as a major strand in the GPS standards, and has resulted in becoming a leader in statistics education in K-12 (Franklin).

For example, consider the course Algebra 1 and the following statistics and probability standards in QCC and its revision for GPS assigned in Table 1.1. From this, it can be seen that the GPS standard has become more descriptive and specific for a clearer understanding. In this example, the QCC Algebra 1 statistics topics have been placed in Grade 7 GPS. This indicates that the breadth of the grades 9-12 QCC standards have been dispersed among different grade levels in GPS, including earlier grades such as 6-8. Also, since previous QCC standards found in grades 9-12 are present in grade 7 math GPS, then it can be concluded that the GPS standards add rigor to the curriculum.

Table 1.1 Algebra 1 QCC vs Statistics and Probability GPS

QCC Standard	GPS Standard
35: Topic: Statistics Summarize data in various ways, including mean, median, mode, and range	MCC7.SP.4 Use measures of center and measures of variability for numerical data from random samples to draw informal comparative inferences about two populations. (Grade 7)
36: Topic: Probability Identifies possible outcomes of simple experiments and predicts or describes the probability of a given event expressed as a rational number from 0 to 1.	MCC7.SP.5 Understand that the probability of a chance event is a number between 0 and 1 that expresses the likelihood of the event occurring. (Grade 7)
37: Topic: Probability Conducts and interprets a compound probability experiment.	MCC7.SP.8 Find probabilities of compound events using organized lists, tables, tree diagrams, and simulation. (Grade 7)

Most recently, in 2010 the Common Core math standards were created, and adopted by Georgia to create the current CCGPS math. According to the Achieve Common Core Comparison Tool, 90% of the overall GPS is aligned with the new Common Core math standards. Therefore, Georgia's GPS was already very close to meeting the Common Core math standards (Georgia State Board). Since the Common Core math standards for high school were not divided into courses, rather just by topics for grades 9-12, it was the respective state's decision on how to place the standards accordingly into their courses. Georgia's approach to this was to keep four main math courses – Coordinate Algebra, Analytic Geometry, Advanced Algebra, and Pre-calculus. Along with these courses are a growing number of elective options for students, such as Advanced Mathematical Decision Making, Mathematics of Finance, and non-AP Calculus. These courses are growing as courses are currently being created, one example would be the course Statistical Reasoning.



The CCGPS Math standards were divided by conceptual categories for grades 9-12. Consider one of the conceptual categories in CCGPS - Statistics and Probability, divided into three subcategories. These three subcategories of the Statistics and Probability standard clusters are found in CCGPS. Tables 1.2 to 1.4 show the development of these standards from the QCC, and GPS. Each respective CCGPS standard is traced by showing where it can be found in the previous standards. It can be seen from the tables that some standards are not specifically given in one set of standards but were presented in previous or later standards. Since the standards became much more detailed and specific as it developed to the CCGPS some standards do not seem to be present in previous standards.

Table 1.2 shows the focus on the use of context of data and using technology in the CCGPS. Fitting a function to the data and analyzing its residuals are standards given in CCGPS that were not explicitly stated in GPS and QCC. The use of technology in interpretation of the correlation coefficient, and distinguishing between correlation and causation are not found in QCC but are in both GPS and CCGPS. In general, QCC lacked emphasis on analyzing data that is imperative for students to be able to work with real-life data presented to them in their everyday life. Both the GPS and CCGPS include this; however the CCGPS also states that the students are to interpret the slope and intercept of their linear model in the context of the data. The relation back to the context of the data is one of the more prominent difference between GPS and CCGPS.

Table 1.2 Interpreting Categorical and Quantitative Data

Standard	CCGPS	GPS	QCC
Represent quantitative and categorical data with plots	Coordinate Algebra	Math 2	Algebra 3
Compare center and spread of data sets without using standard deviation and accounting for outliers	Coordinate Algebra	Math 1	Algebra 1
Summarize categorical data in two-way frequency tables	Coordinate Algebra		Algebra 3
Represent data on two quantitative variables on a scatter plot	Coordinate Algebra	Math 2	
Fit a function to the data emphasizing linear and exponential model	Coordinate Algebra		Advanced Algebra and Trigonometry
Informally assess fit by analyzing residuals	Coordinate Algebra		
Interpret slope and intercept of linear model in context of the data	Coordinate Algebra		
Compute using technology and interpret the correlation coefficient of a linear fit	Coordinate Algebra	Math 2	
Distinguish between correlation and causation	Coordinate Algebra	Math 2	
Fit a function to data on two quantitative variables emphasizing quadratic model	Analytic Geometry	Math 2	Advanced Algebra and Trigonometry
Compare center and spread of data sets including using standard deviation	Advanced Algebra	Math 2	Advanced Algebra and Trigonometry, Algebra 3
Use mean and standard deviation to fit to a normal distribution and to estimate population percentages	Advanced Algebra	Math 2	Advanced Algebra and Trigonometry, Algebra 3

Table 1.3 shows that in QCC, GPS, and CCGPS independence of events are addressed. The conditional probability of two events and their independence is covered in all three standards. However, constructing and interpreting two-way frequency tables is only introduced in the CCGPS. The standards for probability were not very extensive in the QCC and GPS standards, and are still relatively less than the standards given for interpreting data.

Table 1.3 Conditional Probability and the Rules of Probability

Standard	CCGPS	GPS	QCC
Describe events as subsets of a sample space using characteristics of the outcomes	Analytic Geometry		Advanced Algebra and Trigonometry, Algebra 3
Understand and determine independence of two events	Analytic Geometry	Math 1	
Understand conditional probability of two events and determine independence of each event	Analytic Geometry	Math 1	Algebra 3
Construct and interpret two-way frequency tables and determine if events are independent	Analytic Geometry		
Compute probabilities of compound events in a uniform probability model	Analytic Geometry		Algebra 1

Table 1.4 presents standards covered in the advanced algebra course under CCGPS. These standards cover topics that involve large data sets from random samples of populations, surveys, and experiments. Students should be able to make inferences about these data sets. In QCC this is only covered in a course called Statistics, which was not one of the required common courses required, thus most students never received exposure to these topics. In GPS, they are minimally addressed in Math III or IV; again the chance of students covering these standards was unlikely. However, in the CCGPS these standards are very specific and being in Advanced Algebra, the third year course, the students will be exposed to these topics.

Table 1.4 Making Inferences and Justifying Conclusions

Standard			
Make inferences about population parameters based on a random sample from the population	Advanced Algebra	Math 1	Statistics
Decide if a specified model is consistent with results from a given data-generating process	Advanced Algebra		Statistics
Recognize differences among sample surveys, experiments, and observational studies explaining how randomization relates to each	Advanced Algebra	Math 3	
Use data from a sample survey to estimate a population mean or proportion, develop a margin of error through simulation models for random sampling	Advanced Algebra	Math 4	Statistics
Use data from a randomized experiment to compare two treatments, use simulations to decide if differences between parameters are significant	Advanced Algebra		
Evaluate reports based on data	Advanced Algebra		Concepts of Probability and Statistics

### 1.3 Motivation for Revision of Lesson Plans

In order for our citizens to be quantitatively literate, we must prepare students to make informed decisions by presenting problems in real-world contexts which students can relate to, and care about. According to Bernard Madison, professor of mathematics at University of Arkansas, “although high school and introductory college mathematics do include some so-called real-world problems, these very often are not embedded in the world of any student” (Madison). From Madison’s statement, we consider the context in which we present problems to students is not a matter of using real data sets, rather it is imperative that the context be one to which students can relate in their own personal life. Through contextual problems, students will understand the concepts and processes used

in analyzing and interpreting data. According to Richard Scheaffer, professor emeritus of statistics at University of Florida, “in addition to being important in their own right, data analysis skills help build connections between mathematics and other subjects in the school curriculum and to the world outside of the classroom. Such connections are essential if the students are to comprehend the importance of mathematics and the role it plays in virtually all aspects of life” (Scheaffer, 1990).

There are two things to consider when discussing the difference between mathematical thinking and statistical thinking. “In mathematics, context obscures the mathematical structure. But in statistics and data analysis, context provides the meaning” (Franklin). Scheaffer states that, “although statistics uses mathematics, the key to statistical thinking is the context of a real problem and how data might be collected and analyzed to help solve that problem” (Scheaffer, 2003). Design of surveys and experiments plays an important role in modern science, such as demographic surveys of the Census Bureau and economic surveys of the Bureau of Labor Statistics, as well as experiments in healthcare. Statistics emphasizes the context, and uses mathematics as one of the main tools for practical problem solving (Scheaffer, 2003).

The lack of teacher preparation is one of the main factors of poor statistical teaching in the classroom. It has been recommended for many years that to improve math and science education, there must be stronger teacher preparation (National Commission). There is a need for extensive training in teaching mathematics and statistics in context. This is especially true for statistics, as many mathematics graduates lack the statistics education preparation due to statistics courses rarely being required as a course in general education programs. This weak training in statistics leaves teachers unprepared to teach the data analysis and probability present in the school curriculum (Madison). In recent years, there has been an increase in the focus of preparing future teachers, proving more clearly the belief that teachers will be mathematically well-prepared if they receive a degree in mathematics. This ignores the importance of pedagogy, the need for teachers to learn strategies for creating lessons and tasks that will help students gain understanding of mathematics. Teachers who understand higher mathematics often don’t have the training necessary to have an idea on how to translate higher math knowledge into simpler form suitable for students (Steen).

Also, teachers tend to teach according to the test, given the lack of a sufficient amount of time to cover all standards given. This result in choosing to teach only the topics covered in the End of Course Test (EOCT). Many times this means statistical concepts are never covered. In addition, there was no EOCT for Math III and IV, therefore many teachers chose to teach only the topics they were comfortable in teaching. This resulted in some teachers skipping the statistics unit of the course. (Humphrey)

The Georgia Department of Education provided lesson plans for teachers to use in their classroom that were determined to be aligned with CCGPS. Lesson plans were provided for Unit 4: Describing Data in the Coordinate Algebra course, and Unit 7: Applications of Probability in the Analytic Geometry course. In choosing lesson plans to revise, NCTM standards and the GAISE K-12 Report were considered. Both

frameworks consider linear models as an important statistical topic for high school students to learn.

Applications of linear models are seen in other courses such as social science, and physical science. Linear models are sometimes called trend lines, representing trends over a period of time. These trends can represent consumer prices or stocks. Linear models are seen throughout the news, in finance and economics. We considered the vast amount of applications and the importance of the topic of linear models in choosing lesson plans to revise. For these reasons, the learning tasks and performance task given in Unit 4: Describing Data in the Coordinate Algebra course covering linear models were chosen for revision.

## CHAPTER 2

### LEARNING TASK 1: SIMPLE LINEAR REGRESSION

The original learning task, Spaghetti Regression, found in Appendix A has the following mathematical goal, “to investigate the concept of goodness of fit and develop an understanding of residuals in determining a line of best-fit”. The revised learning task follows the same goal, and in addition students should be able to define a residual, and know how it is measured. They should also be able to determine a best-fit line based on the residual measure.

Figure 2.2 Overhead 1 in the original lesson served to introduce the topic of goodness-of-fit by discussing with the students why the line on the top graph fits the data better than the bottom graph. However, the graphs did not show clearly a best-fit line, and the bottom graph did not have a line fit to the data points. To address this, the revised lesson gives a new overhead to use as an introduction to the lesson. The new overhead 1 is an example of a scatter plot of data representing years of education versus income per year. Figure 2.3 Overhead 2 is the same example, but with the best-fit line included on the graph.

The introduction of residuals has been revised to define to the students what a residual is and how is it measured. This creates an understanding of a residual, rather than having the students use a piece of spaghetti to find the regression and having some incorrectly measure their distances by having them measure in different directions as proposed in the original lesson (contrary to the statistical definition of a residual).

The original task does not address the notion of sum of squared residuals. This method of determining the best-fit line by minimizing the sum of squared residuals is missing in the original learning task. Students are directed to line up their spaghetti pieces to see whose line fits best by determining whose spaghetti pieces ended up shortest. However, this allows for multiple best-fit lines, which is addressed in the revised task. In this task, students are to calculate the residual measure of the best-fit line they created, by finding the sum of squared residuals. Also, a class discussion is presented and the teacher is provided with information to address some possible questions students may ask. This better prepares teachers to be able to teach the topic of regression with a deeper understanding.

The worksheet provided for students has been revised in such a way that there is context in the data provided. The original scatter plot does not have axis labels, nor does it have a scale for which the students can determine the distance of each point. In the revised task, data from TV watching hours versus test scores is given on a scatter plot for students to eye-ball a best-fit line, and measure their residual sum of squares. Students are then asked to make conclusions of their findings in terms of the context given.

## 2.1 Teaching Guide

### Learning Task 1: Simple Linear Regression

#### Mathematical Goals

- To investigate the concept of best-fit and develop an understanding of residuals in determining a line of best-fit

Table 2.1 Learning Task 1 Standards

Common Core State Standards	Student Worksheet Questions
<b>MCC9-12.S.ID.6</b> Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.	1, 2
<b>MCC9-12.S.ID.6b</b> Informally assess the fit of a function by plotting and analyzing residuals.	3, 4, 5, 9
<b>MCC9-12.S.ID.6c</b> Fit a linear function for a scatter plot that suggests a linear association.	6, 7, 8

#### Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.

#### Introduction

Students will investigate the concept of the “goodness-of-fit” by determining the regression line or best-fit line for a set of data. This is the first exploration in a series of three activities to explore a best-fit line and residuals. Fitting of an equation to the graph of a data set is covered in all mathematics courses from Algebra to Calculus and beyond. The objective of this activity is to explore the concept in-depth.

In real life, functions arise from data gathered through observations or experiments. These data rarely fall neatly into a straight line or along a curve. There is variability in real data, and it is up to the student to find the function that best 'fits' the data. Regression, in its many facets, is probably the most widely used statistical methodology in existence. It is the basis of almost all modeling.

Students create scatter plots to develop an understanding of the relationships of bivariate data; this includes studying correlations and creating models from which they will predict and make critical judgments. As always, it is beneficial for students to generate their own data. This gives them ownership of the data and gives them insight



into the process of collecting reliable data. Teachers should naturally encourage the students to discuss important concepts such as goodness-of fit. Using the graphing calculator facilitates this understanding. Students will be curious about how the linear functions are created, and this activity should help students develop this understanding.

**Materials:**

- Transparencies of Overhead 1 and 2, and Measuring Notes
- Handouts – copy for each student of the Scatter plot,
- Student Activity: Simple Linear Regression, and Measuring Notes
- Rulers

**Procedure:**

- I. Introduce the topic of Goodness-of-Fit with Overhead 1 and 2.
  - a. Define a scatterplot as a graphical representation of data points plotted on a graph from a data set with  $x$  (predictor) and  $y$  (response) variables.
  - b. Figure 2.2 Overhead 1 shows a scatterplot representing a data set of Number of Years of Education ( $x$ ) and Average Income ( $y$ ) from a sample of ten people. Discuss with the students that when data are collected from observations or experiments, the data rarely fits nicely on a straight line or curve. Many times data “scattered” on a graph will be encountered due to variability in real data. Ask the students the following questions:
    - i. Do you think there is a relationship between the years of education with the amount of income?
    - ii. If you want to have an idea of how much your income will be after you attended college for 4 years or if you didn’t go to college at all, how can you use the graph to estimate your income?
  - c. Figure 2.3 Overhead 2 shows the same scatterplot with the addition of the best-fit line drawn on the graph as well. Ask the students the following questions:
    - i. How can we determine that this is the best fit line?
    - ii. How can we find the equation for this line?
- II. Introduce the concept of residual and how it is measured using the Measuring Notes handout. Discuss that the residuals are used to measure the goodness of fit of the best-fit line and ask the students how they can use the residuals to find the best fit line. From this, the class should come to a consensus that the residuals measure the distance from the actual value of the data point to the predicted value on the line. Direct the class to infer that the best-fit line will have the minimum error, thus the minimum residual sum.

*\*It is important to note here that though the class makes this inference regarding the residual sum, it is imperative that students are corrected after*

*the Learning Task that using the least squares method, we are to minimize the sum of squared residuals.\**

- III. Given the following data set, have each student create a scatter plot of the data on their empty graph, Figure 2.7.

Table 2.2 TV Time & Test Score

TV Time (Hours)	Test Score
30	70
12	85
30	75
20	85
10	100
20	88
15	85
12	90
15	90
11	90
16	95
20	85
19	85

- IV. Now the students all have a scatterplot of Time Spent Watching TV (predictor) and Test Score (response). Have the students draw what they believe to be the best-fit line by eyeballing and have them measure the residuals.
- V. Have the students answer the questions on their Learning Task Worksheet.
- VI. Discuss with the class who has the best-fit line based on their previous inference that they must minimize the residual sum. Now, question their inference by using one of the students' best-fit line and asking what happens with residuals that are positive and negative when you add them together. Ask the students how this affects their measure for the goodness of fit.

Guide the students to conclude that when two data points have an equal distance from the line but one is above the line and one is below the line then the residuals cancel each other out, therefore seeming like there is no error. This can be shown through the example data set of Years of Education vs Income, where we give an example of residual approximately -3, and a residual of approximately 3. Ask the students to suppose all the data points paired up in this manner and cancelled each other, in that case we have sum

of residuals as zero; however it is clear to see that we have a positive correlation and error in the model.

Probe the class for suggestions for how to correct this problem. Most likely the students will propose the idea of using absolute value. Though this sounds like a great solution, demonstrate to the students that this could lead to multiple best-fit line equations. Because of this lack of uniqueness we search for a better solution. Reach a consensus that by taking the square of each residual there will no longer be negative residual values, yet produces one best fit line. So instead of taking the sum of the residuals, we take the sum of the squared residuals to measure the goodness of fit.

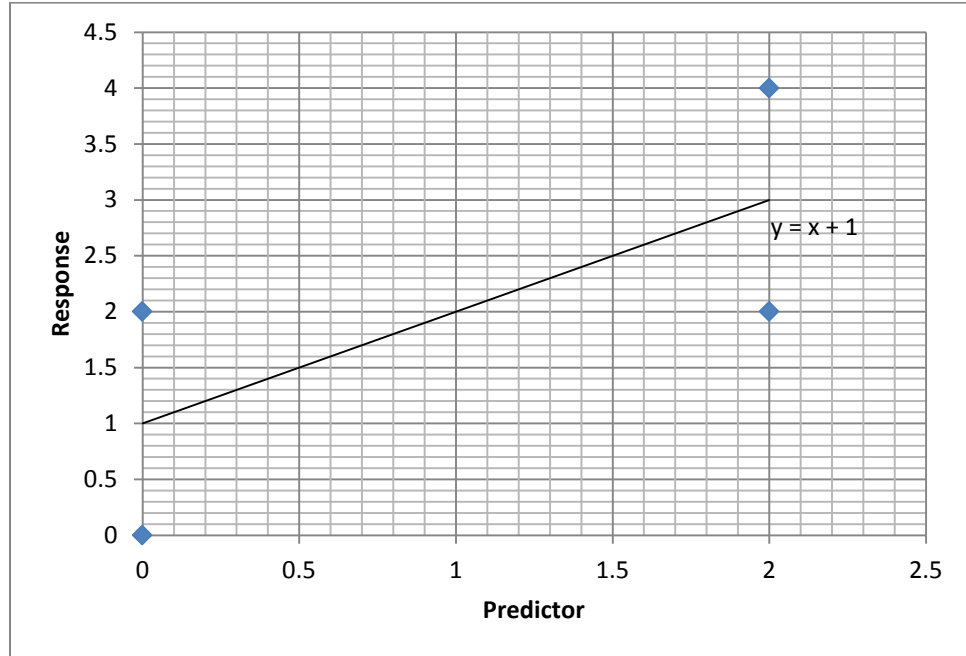
*\*An example of this can be given with the following.\**

Graph the points (0,0), (0,2), (2,2) and (2,4). Then ask the students what the best-fit line is by finding the absolute value of the residuals. The students should agree upon  $y = 1 + x$  as the best-fit line. Now point out to the students that other equations such as  $y = 2$ ,  $y = x$ , and  $y = 0.2 + 1.6x$  also all have absolute value of residuals equal to 4. However, consider their sum of squared residuals and notice that the sum of squared residuals for these lines are larger than 4, which is the sum of squared residuals for  $y = 1 + x$ . This shows that by finding the minimum sum of squared residuals, we are able to find a unique best-fit line. On the other hand, the sum of the absolute value of residuals does not give us a unique best-fit line (Watkins).

Table 2.3 Measuring Residuals

Best-Fit Line Equation	Absolute Value Residuals	Sum of Squared Residuals
$y = 1 + x$	4	4
$y = 2$	4	8
$y = x$	4	8
$Y = 0.2 + 1.6x$	4	5.6

Figure 2.1 Best Fit Line Example



- VII. Have the students square all of their residuals and find the sum. Now, ask the students to share their sum of squared residuals to find who has the best fit line from the class.
- VIII. As a class determine the equation of the best-fit line using the point-slope form.
- IX. Using the best-fit line equation from the class, ask the students how they can use the line to predict what their test score will be based on how many hours of TV they watch per week.
- X. Ask the students what they can conclude about the relationship between the number of hours of TV they watch per week and their test score. Be sure they use context in interpreting the numerical value of the slope.
- XI. For deeper thinking, an extension activity can be given to the students by discussing the appropriateness of the linear regression model using residual plots. A residual plot analysis information sheet, and calculator instructions are provided for reference.

Figure 2.2 Overhead 1

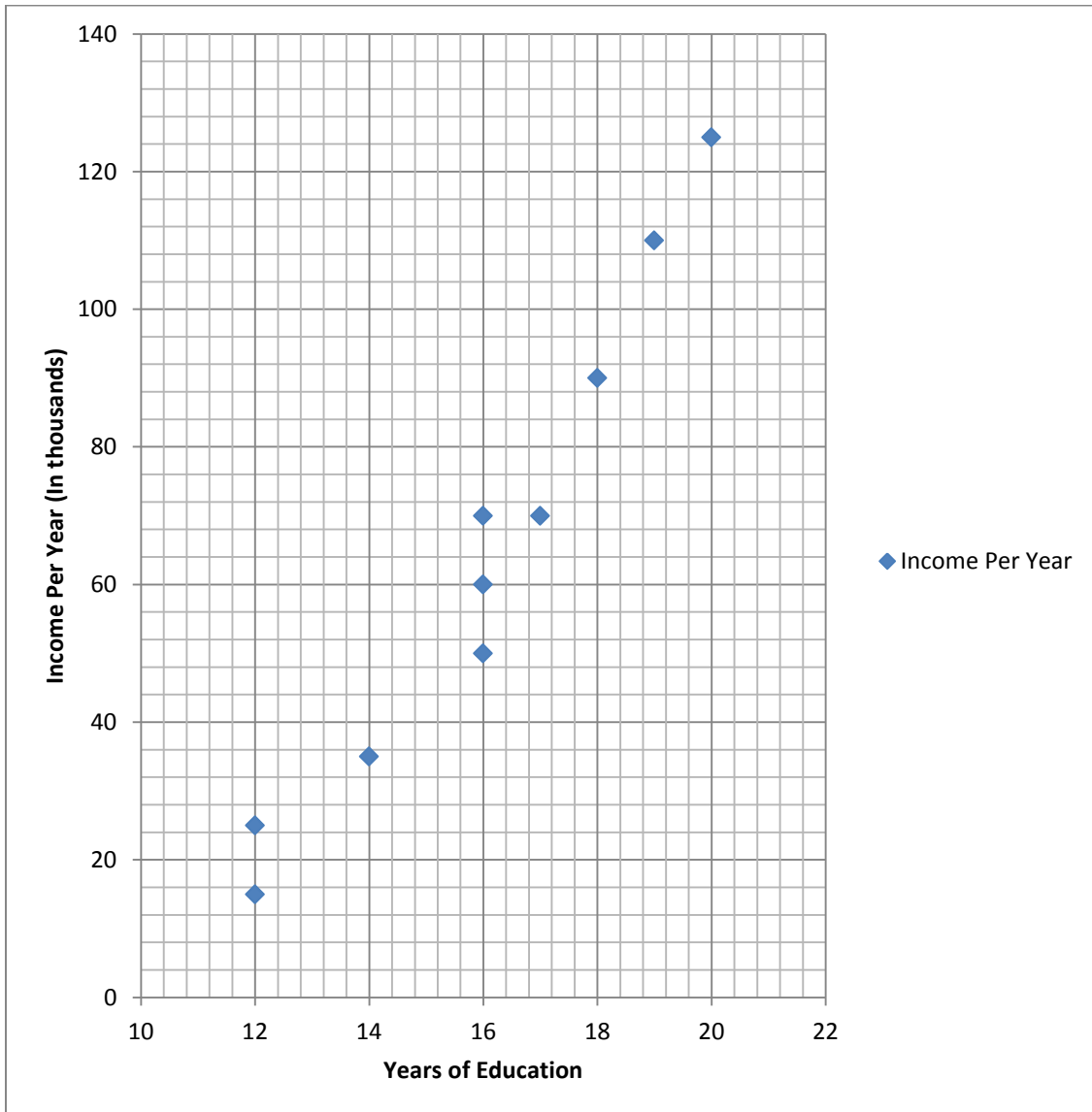


Figure 2.3 Overhead 2

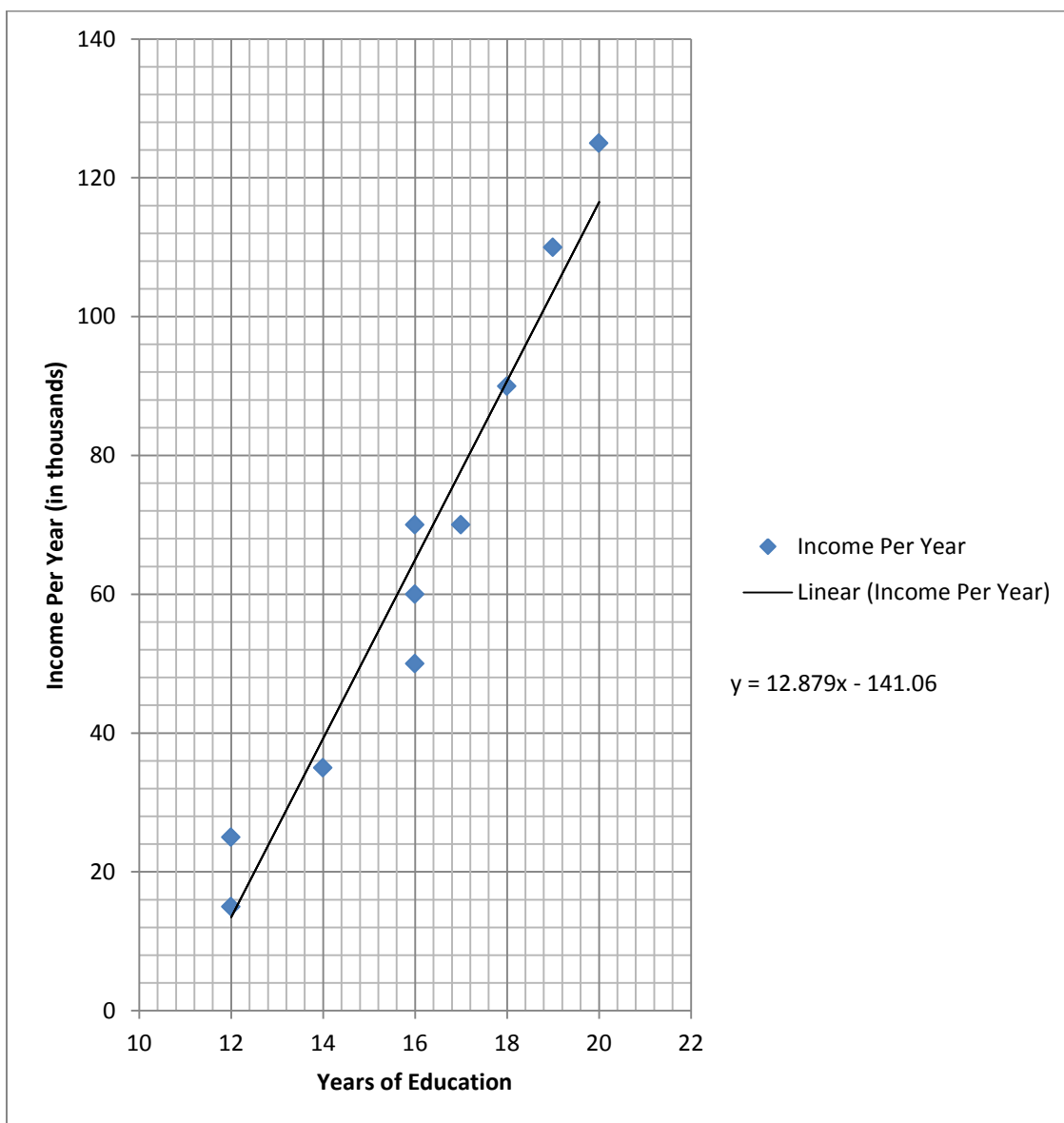
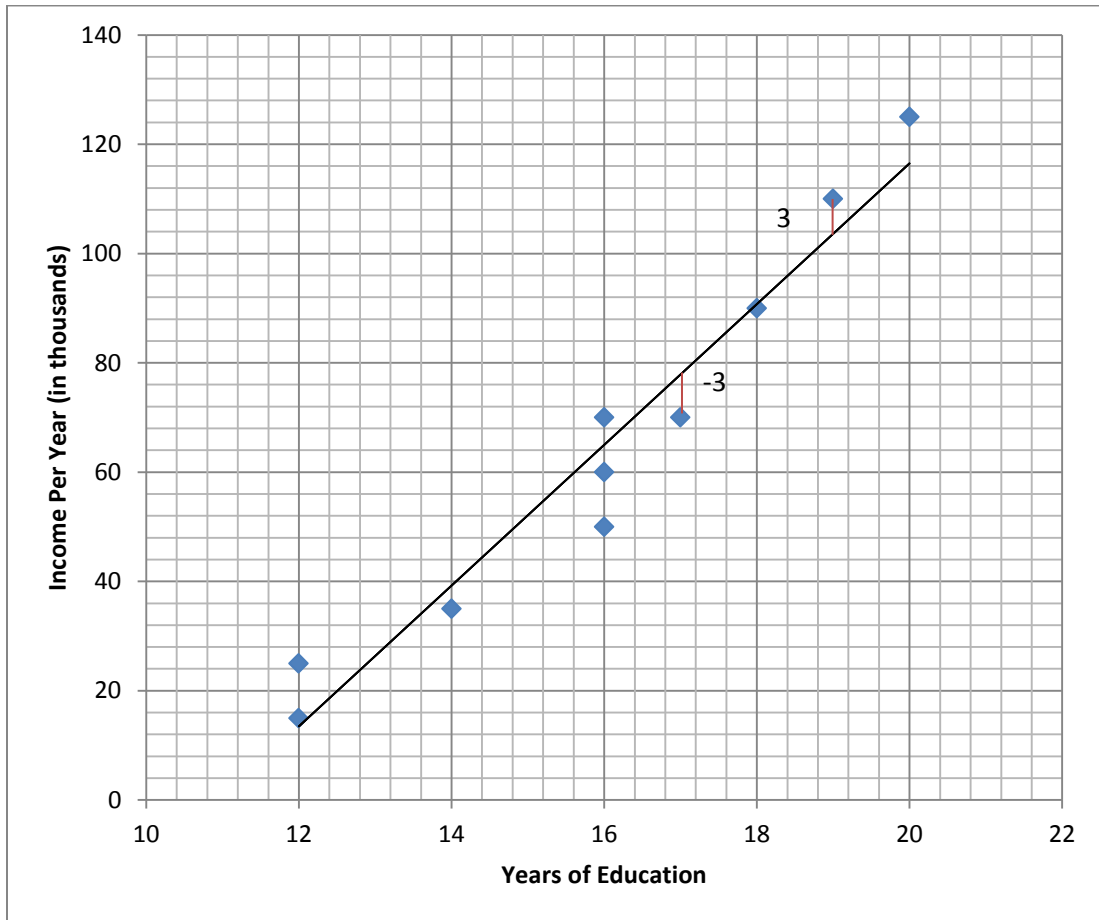


Figure 2.4 Measuring Residual Distances

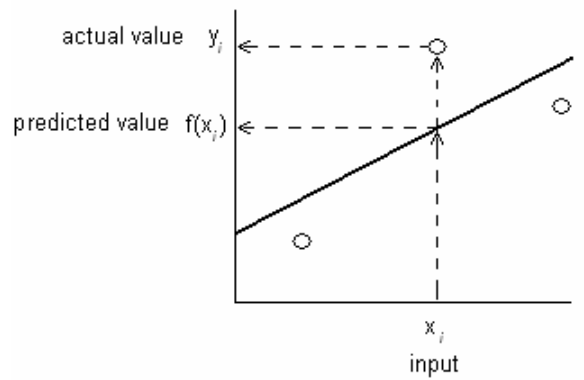


**Teaching Note:**

While discussing the definition of residual, discuss the issue that arises when two residuals are of the same absolute value but have different signs by demonstrating the example above. If we take just the value itself in calculating the residual sum, then we could potentially have a residual of zero when all values cancel out, however, this could occur even when there is in fact a linear relationship. To solve this problem, it seems logical then to take the absolute value of each of the residuals, and use their sum to find the best fit line. However, this isn't the best solution to this problem since this does not produce a unique best fit line. This can cause a problem as there could be multiple best fit line equations. This also brings difficulty mathematically as a formula containing absolute value is difficult to minimize using differentiation in Calculus. For these reasons, a preferred solution is to square the residuals to make all the residuals a positive number. Therefore, the minimum sum of squared residuals gives us the unique "best fit" linear equation line.

## Measuring Notes

Figure 2.5 Residual Definition

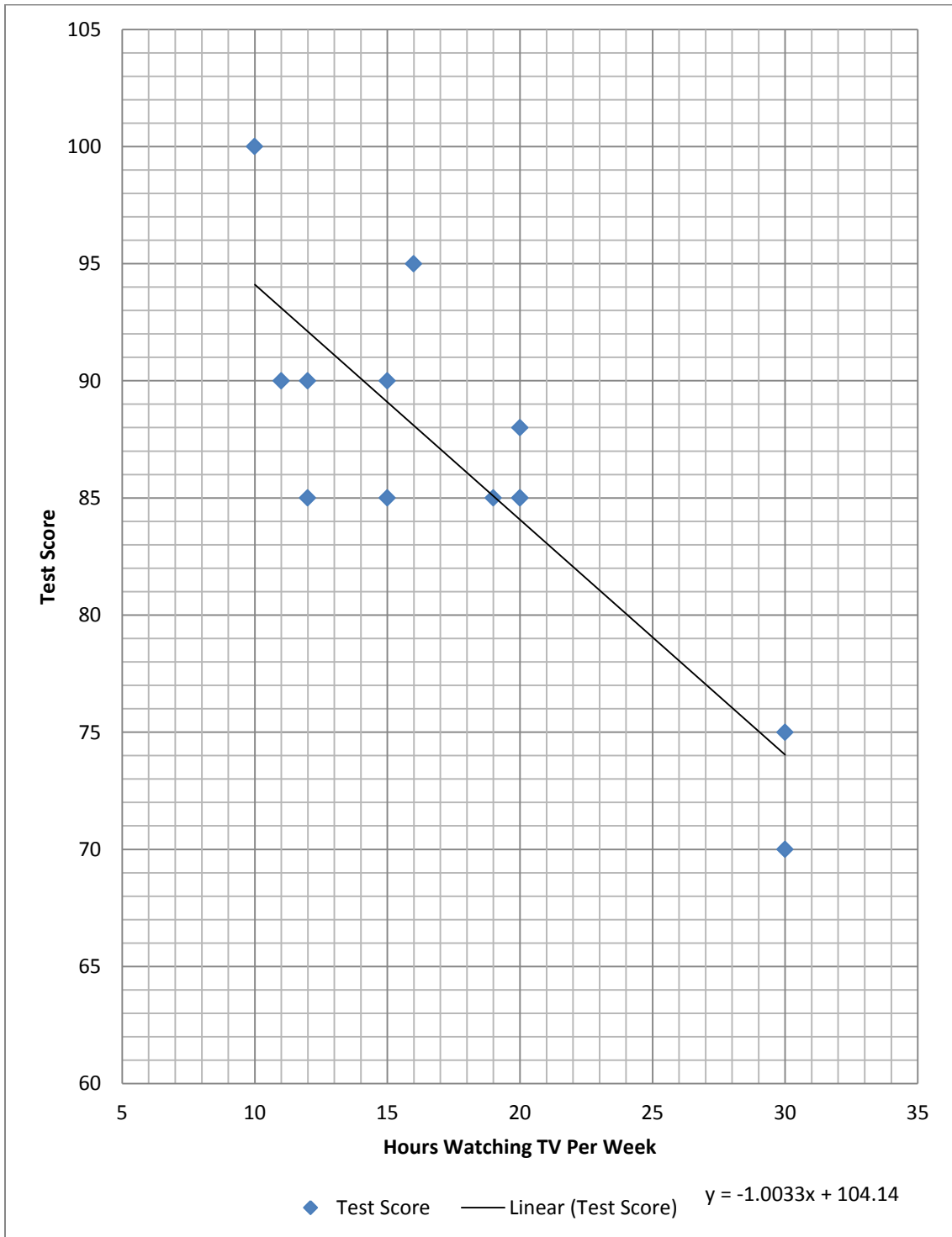


The purpose of regression is to find a function that can model a data set. The function is then used to *predict* the y values (outputs or  $f(x)$  for any given input  $x$ . So, the vertical distance represents how far off the prediction is from the actual data point (i.e., the “error” in each prediction.) Residuals are calculated by subtracting the model’s predicted values,  $f(x_i)$ , from the observed values,  $y_i$ .

$$\text{Residual} = y_i - f(x_i)$$



Figure 2.6 TV Hours vs Test Score with Best-Fit Line



### Teaching Extension: Residual Plots

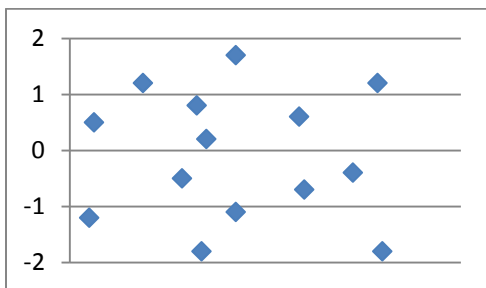
A residual plot plots the residual ( $y_i - f(x_i)$ ) against the explanatory variable  $x$ . This shows the appropriateness of the linear regression equation for the data set.

The residual plot should not have any pattern, rather the residuals should be randomly distributed around zero. A residual plot with a random pattern represents a good fit for the linear model. Residual plots with non-random patterns indicate that perhaps a higher order model is more appropriate, or a transformation of the data is suggested.

The following are examples of common residual plots.

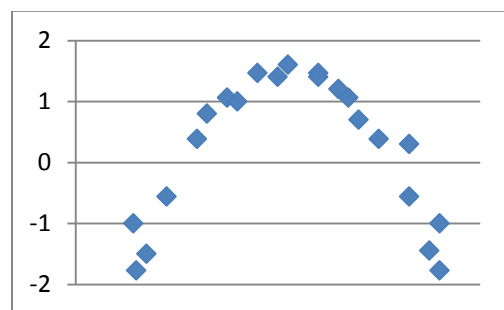
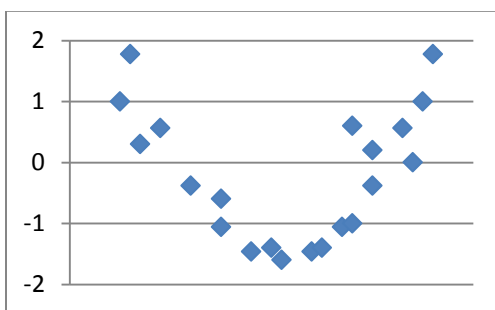
#### Random Pattern (Good fit for a linear model):

Figure 2.7 Random Residual Plot



#### Non-Random Pattern:

Figure 2.8 Non-Random Residual Plot



### Plotting Residual Plots using a TI-83/84 Calculator

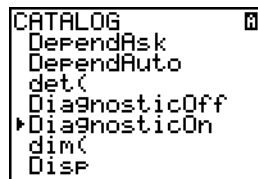
We will use the following example data set to introduce how to use a graphing calculator to plot the residual plot of the linear regression equation.

Table 2.3 Education and Income Data

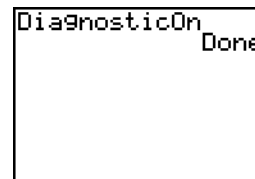
Years of Education	Income Per Year (in thousands)
19	110
20	125
16	60
16	70
18	90
12	15
14	35
12	25
16	50
17	70

1. Before entering our data set into the calculator, first turn on the diagnostics so our correlation coefficient  $r$  will be shown. Do this by pressing  $\boxed{2\text{nd}}$   $\boxed{[\text{CATALOG}]}$  then press  $\boxed{x^{-1}}$  then scrolling down to DiagnosticOn and pressing  $\boxed{\text{ENTER}}$  twice.

Figure 2.9 Calculator Diagnostic



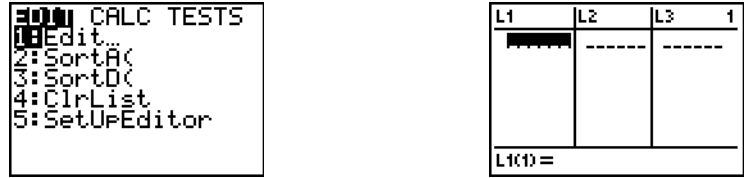
```
CATALOG
DependAsk
DependAuto
det(
DiagnosticOff
DiagnosticOn
dim(
Disp
```



```
DiagnosticOn
Done
```

2. Enter the data set into the calculator. Do this by pressing  $\boxed{\text{STAT}}$  and press  $\boxed{\text{ENTER}}$  when Edit... is highlighted. This will take you to the lists screen. L1, L2, L3 are different lists of data, in our case we will put our Years of Education data in L1 and Income in L2.

Figure 2.10 Calculator Listing



- To clear the data lists, press **[STAT]** and press **[ENTER]** when Edit... is highlighted. This will take you to the lists screen. To clear a list, use the up arrow to place the cursor on the list name then press **[CLEAR]** and **[ENTER]**. Take note that pressing **[DEL]** instead of **[CLEAR]** will delete the list from your calculator rather than clearing the data in the list.
- Enter the data Years of Education into L1 then right arrow to the next column and enter the Income data into L2. After, press **[2nd]** **[MODE]** to quit the listing screen.  
*Note: The lists must be the same length (L1 and L2 should have the same number of data elements in each). If they are not the same length then, ERR: DIM MISMATCH will be displayed when attempting to graph the data or perform the regression.*

Figure 2.11 Data Lists

L1	L2	L3	1
20	110		
16	125	-----	
16	80		
16	70		
18	90		
12	15		
14	35		
L1()=19			

- To plot the residual plot, we will need to find the linear regression. To do this press **[STAT]** and scroll to the right to CALC. Press **[4]** to choose LinReg(ax+b) then specify the lists in which the explanatory variable, and response variable are stored in, do this by pressing **[STAT]** **[L1]** **[,]** **[L2]** **[,]**. Then to save the regression line as a function (perhaps to show it with the original data), press **[VARS]** **[▶]** to select Y-VARS, then select 1: Function and press **[ENTER]**, then select where you want to store the function, such as Y<sub>1</sub> then press **[ENTER]** and the screen should look like the one below in Figure 2.11.

Figure 2.12 Linear Regression



6. Before creating the residual plot, we turn off other plots, so that we only see the residual plot. First, check to make sure that any functions are not plotted. To do this, press  $\boxed{Y=}$  and check that all equal signs are not bold. If an equal sign is bold, move the cursor to the equal sign and press  $\boxed{\text{ENTER}}$ , this should change the equal sign to remove the bold feature. Also, ensure that all other statplots are off by pressing  $\boxed{2\text{nd}} \boxed{Y=}$ , this will give you the STAT PLOTS menu, where you check that all plots are off.
7. Now, to generate the residual plot press  $\boxed{2\text{nd}} \boxed{Y=}$ , this will give you the STAT PLOTS menu. Turn on and define Plot 1 by highlighting 1 then pressing  $\boxed{\text{ENTER}}$ . This will bring you to the Plot1 screen. Highlight On and press  $\boxed{\text{ENTER}}$ . The Xlist should have the list where the explanatory variable was stored, in our case L1. The Ylist should have the residuals list. To state RESID on the Ylist, press  $\boxed{2\text{nd}} \boxed{\text{STAT}} \boxed{\text{LIST}}$  and select 7:RESID. Press  $\boxed{2\text{nd}} \boxed{\text{MODE}}$  to quit the current screen.

Figure 2.13 Stat Plots



8. Now, to view the residual press  $\boxed{\text{ZOOM}} \boxed{9}$  or scroll down to 9: ZoomStat and press  $\boxed{\text{ENTER}}$ . Have the students observe the residual plot and determine if the linear regression equation is appropriate for the given data set.

Figure 2.14 Zoom Stat



## 2.2 Student Worksheet: Simple Linear Regression

Name \_\_\_\_\_ Date \_\_\_\_\_

### Learning Task: Simple Linear Regression

#### **Common Core State Standards for Mathematical Practice**

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.

#### **Common Core State Standards**

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

#### **Part I**

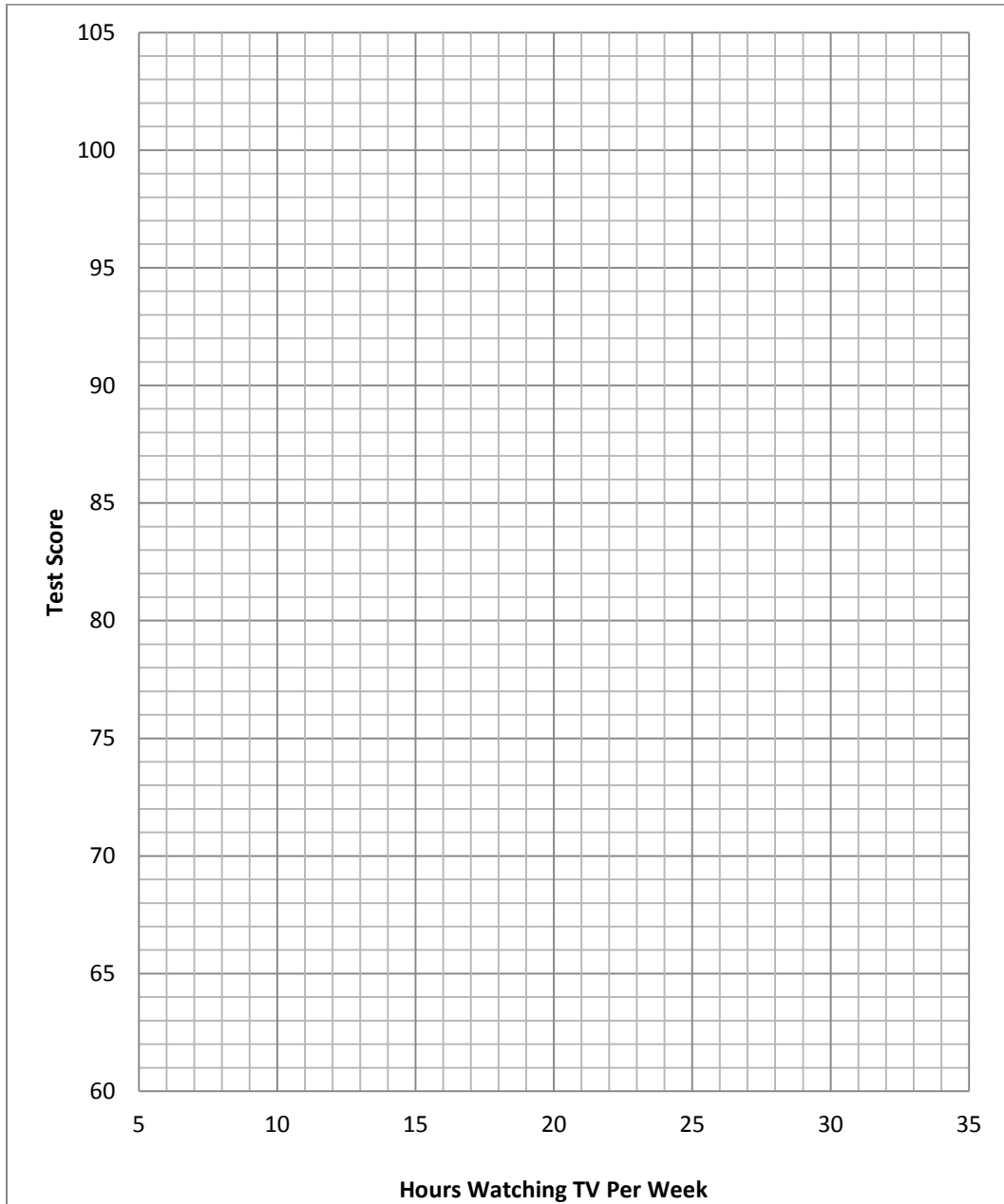
1. Create a scatter plot of the following data set TV Time and Test Score. Plot the data points on the graph provided.

Table 2.4 TV Time and Test Score

TV Time (Hours)	Test Score
30	70
12	85
30	75
20	85
10	100
20	88
15	85
12	90
15	90
11	90
16	95
20	85
19	85

2. Examine the scatter plot you created and visually determine a line of best-fit (or trend line). Draw your best-fit line on your scatter plot.

Figure 2.15 Hours of TV vs Test Score



3. Now investigate the “goodness” of the fit. Measure the residual using the method from the Measuring Notes sheet. Repeat this for each point in the scatter plot.

*Answers will vary.*

4. Calculate the sum of your residuals.

Total error = \_\_\_\_\_

\* Class discussion before moving to Part II. \*

### **Part II**

5. Calculate the square of each residual. Find the sum of your squared residuals.

*Answers will vary. Each residual found in question 3 is to be squared.*

Total residual sum of squares = \_\_\_\_\_

6. As a class find the equation of the best-fit line.

*The best-fit line found using the calculator is  $y = -1.00x + 104.14$*

***Figure 2.16 Linear Regression Output***

```
LinReg
y=ax+b
a=-1.003339405
b=104.1360049
r2=.7194698923
r=-.8482157109
```



7. Using the class best-fit line, what do you think your test score will be based on how many hours of TV you watch per week?

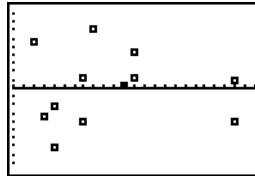
*Answers will vary depending on how many hours of TV the student thinks he/she watches per week.*

8. What can you conclude about the relationship between the number of hours of TV students watch per week and your test score?

*The students should conclude that for every hour of TV they spend watching per week, their test score decreases by about 1 point.*

9. **Extension:** Plot the residual plot of the data using your calculator. What can you conclude about the linear model for the data set? Is the linear model an appropriate representation of the data?

*Figure 2.17 Hours of TV vs Test Score Residual Plot*



*The linear model is appropriate for this data set, as the residual plot appear to be randomly distributed around the x-axis.*

## 2.3 Comparison Summary

In considering the original learning task, some shortcomings were found. One of the shortcomings was that it stated three standards to be addressed in the lesson – represent the data on a scatter plot and describe how the variables are related, informally assess the fit of the function by defining and analyzing residuals, and fit a linear function that suggest a linear association.

The first standard states that students are to represent the data on a scatter plot and describe how the variables are related. The students were given a scatter plot, so they did not need to represent the data on a scatter plot themselves. There was also no question asking the students to describe the relation of the two variables. Also in this case, there was no context for the scatter plot data points, so the students could not relate the two points with meaning. The revised task asks the students to create a scatter plot given a data set, and asks the students to make a conclusion of the association between the two variables.

The students are also to fit a function that suggests a linear association and analyze the residual. In both the original and revised task the students eye-ball a line as their best-fit line, and calculate the residual of their line. However, the original task does not provide the students with the definition of a residual, nor how to properly measure the distance of their fitted line from their data points. The students line up their spaghetti distances and check who has the shortest total distance. This does not teach the students how to correctly calculate a residual, using one of the methods such as least-squares method. The revised task asks the students to find who has the best-fit line in the class by asking the students to calculate their residual by finding their sum of squared residual. The students analyze their residual by comparing them with the rest of the class, and concluding who has the best-fit line by the minimum sum of squared residuals.

In the revised task the students are asked to make a conclusion about the relationship between the number of hours of TV students watch per week and their respective test score.

An extension of the task is to plot the residual plot for the data for checking the appropriateness of the linear model. The second learning task will continue the lesson of linear models by introducing correlation coefficients as a measure of association of two numeric variables.

## CHAPTER 3

### LEARNING TASK 2: TV WATCHING HOURS AND TEST GRADES

The goals of the original learning task TV/Test Grades in Appendix B are to represent data on a scatter plot and describe how the two variables are related, and fit a linear function and assess the fit of the function by plotting and analyzing residuals. The revised task will address similar goals, by having the students represent the data on a scatter plot and fitting a best-fit-line to the data. However, the revised task focuses on the introduction of correlation coefficients and its interpretation. Students should be able to determine if there is a positive, negative, or no correlation between two variables. Students should also understand the difference between correlation and causation.

The original learning task TV/Test Grades in Appendix B does not provide a procedure section for the teachers, only the student worksheet. This lack of procedure section does not provide the teacher the material necessary for teaching the lesson of correlation as the goal of this task. In the revised task, the teacher will define the correlation coefficient of two variables, and discuss how to interpret the value of this coefficient. In the original learning task, this was assumed to have been previously learned; however this topic is not covered in previous courses in the new CCGPS. Students will gain practice in the ability to recognize strong and weak, positive and negative correlations using applets available online for teachers to show students. This will give the students an understanding of interpretation of the correlation coefficient.

Also provided for the teacher is the example data set from the previous task, years of education versus income per year. Using this data set, students will be shown how to make a scatter plot, obtain the best-fit line and correlation coefficient using their graphing calculator. This ensures teachers have the directions needed to direct their students in learning how to compute the correlation coefficient using technology as stated by the standards. Again, we ensure that students relate their findings back to the context by asking probing questions that demand this type of answer.

### 3.1 Teaching Guide

#### **Learning Task 2: TV Watching Hours and Test Grades**

##### **Mathematical Goals**

- Describe how two variables are related using the correlation coefficient
- Use a graphing calculator to draw a scatter plot, fit a linear regression equation, and calculate the correlation coefficient
- Distinguish between correlation and causation of two variables

Table 3.1 Learning Task 2 Standards

<b>Common Core State Standards</b>	<b>Student Worksheet Questions</b>
<b>MCC9-12.S.ID.6</b> Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.	3
<b>MCC9-12.S.ID.6c</b> Fit a linear function for a scatter plot that suggests a linear association. Interpret linear models.	3
<b>MCC9-12.S.ID.7</b> Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.	9
<b>MCC9-12.S.ID.8</b> Compute (using technology) and interpret the correlation coefficient of a linear fit.	3, 4, 5, 6
<b>MCC9-12.S.ID.9</b> Distinguish between correlation and causation.	7, 8

##### **Standards for Mathematical Practice**

- 1. Make sense of problems and persevere in solving them.**
- 2. Reason abstractly and quantitatively.**
- 3. Construct viable arguments and critique the reasoning of others.**
- 4. Model with mathematics.**
- 5. Use appropriate tools strategically.**

##### **Introduction**

Before beginning the task, ask the class what they know about correlation. The correlation coefficient is a measure of how closely two quantitative variables are linearly related, and is a number between  $-1$  and  $1$ . If the values of both variables tend to increase (or if the values of both decrease), the two variables are positively correlated. If one variable tends to decrease as the other increases (or vice versa), the two variables are negatively correlated. If the values of the variables in both sets do not demonstrate a linear relationship, the variables are not correlated. Determining a relationship between two variables, especially from a scatter plot, may be subject to interpretation. The teacher will likely want to have students use a graphing calculator with statistical capabilities to do this task, determining ahead of time which features on the calculator are appropriate.

Lines of good fit may be found using paper-and-pencil techniques (such as writing the equation based on two points) or using a graphing calculator (either generating possible lines to use for guessing and checking or using the regression feature of the calculator to determine a particular function rule). Discuss correlation and causation with the group. Ask them at the end of the task to summarize television watching and test grades and if they believe there is a causal relationship. Have them defend their position based on statistical analysis.

### **Materials**

- pencil
- graphing paper
- graphing calculator or statistical software package

### **Prerequisites**

Students must have knowledge of writing linear equations based on two points and understand correlation.

### **Time Required**

1 to 2 class periods.

### **Procedure:**

1. Recall the previous Learning Task – Simple Linear Regression. Briefly discuss with the students the creation of scatter plots given a set of data, and finding the best-fit line and its equation. Previously, the measure of best-fit was quantified by finding the residual sum of squares. Now, the class will learn how to measure the strength of the linear association between two variables using the correlation coefficient.
2. Using the Correlation Coefficient handout, introduce to the students the definition of correlation coefficient and how it is interpreted.
3. An option for helping students understand correlation, and practice determining positive and negative correlation, show the students several scatter plots available online through correlation applets. You may refer to the Guess the Correlation applet at [www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html](http://www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html) . There are many other applets found online by searching “correlation applets”.
4. Have the students answer Part 1 of their worksheet. From the knowledge they have of correlation, have them predict what they think will be the correlation between different pairs of variables from their data set.
5. Using the data from Income and Education as an example, show the students how to create a scatter plot, fit a line and find the correlation coefficient of their data using their graphing calculator (TI-83/TI-84).

Table 3.2 Education and Income

Years of Education	Income Per Year (in thousands)
19	110
20	125
16	60
16	45
18	85
12	28
14	35
12	24
16	55
17	65

6. Have the students work on Part 2 of their worksheet TV/Test grades.
7. Discuss as a class if the students' instincts were right with their guess of the relationship of the given variables.

## Correlation Coefficient Handout

Correlation coefficients, represented with the letter  $r$ , are used in Statistics to measure how strong a linear relationship is between two variables ( $x$  and  $y$ ). This can be a positive correlation, in which as  $x$  increases,  $y$  also increases, or if  $x$  decreases,  $y$  also decreases. This is a direct relationship, where the behavior of a second variable is the same as the behavior of the first variable. A negative correlation occurs when there is an inverse relationship, where the behavior of a second variable is opposite of the behavior of the first variable. In this case, when  $x$  increases,  $y$  decreases or when  $x$  decreases,  $y$  increases. It is possible that there is no correlation between two variables, that is there is no *linear* relationship between  $x$  and  $y$ . Take note that no correlation does not imply no relationship; we can only conclude we have no linear relationship.

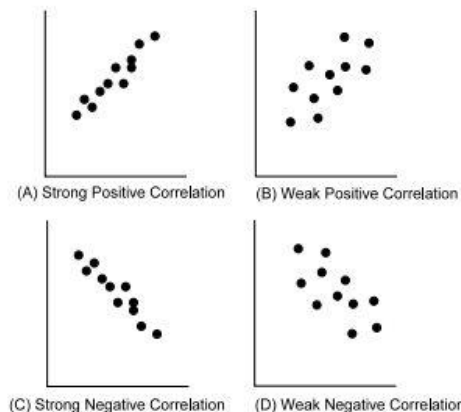
The correlation coefficient can be calculated using the following formula. From the formula, we can see that we are taking the sum of the products of  $x$  and  $y$  relationships, then taking the average by dividing by the number of pairs ( $n-1$ ).

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

The correlation coefficient  $r$ , takes on the value between  $-1$  and  $1$ . The sign of  $r$  defines the direction of the relationship, positive or negative. While  $0$  represents no correlation. To measure the strength of the correlation, we use the following guideline.

- $-1.0 \leq r \leq -0.7$  : Strong Negative Correlation
- $-0.7 < r \leq -0.3$  : Moderate Negative Correlation
- $-0.3 < r < 0$  : Weak Negative Correlation
- $r = 0$  : No Correlation (No *linear* relationship)
- $0 < r < 0.3$  : Weak Positive Correlation
- $0.3 \leq r < 0.7$  : Moderate Positive Correlation
- $0.7 \leq r \leq 1.0$  : Strong Positive Correlation

Figure 3.1 Correlation



\*Note: Although, we almost always consider  $x$  as our predictor variable, and  $y$  our response variable, we cannot make any conclusions on cause and effect. Correlation does not mean causation, we cannot determine if  $x$  causes  $y$  or  $y$  causes  $x$  based on our correlation coefficient.\*

## Teaching Notes:

### Correlation equal zero

When  $r = 0$ , this is interpreted as no correlation. However, assert to students that this does not imply that there is no relationship between the two variables being considered. From this, we can only conclude that there is no *linear* relationship. For example, we could have a quadratic equation, in which case the correlation coefficient can be zero, but this doesn't imply a lack of relationship.

### Association vs. Correlation

Although association and correlation are sometimes used interchangeably, it is important to take note that association is not equivalent to correlation. When we have correlation, we also have association. However, the other case is not necessarily true. If we have association this does not imply we have correlation. Correlation requires a relationship between two *numeric* variables, while association includes the relationship between *categorical* variables. For example, we can find the correlation between years of education and income per year since both variables take on numerical values. On the other hand, we cannot find the correlation between gender and income per year since gender is a categorical variable.

### Association vs. Causation

Measuring association using the correlation coefficient determines the level of strength of the linear relationship between two variables. However, this does not determine the cause of the relationship. In other words, we cannot say that variable  $x$  causes  $y$ , nor can we say variable  $y$  causes  $x$ . Thus, we cannot conclude causation from association. For example, we found that as ice cream sales increase, the rate of drowning deaths also increase. But we cannot conclude that ice cream sales cause drowning deaths, nor can we conclude that drowning deaths cause ice cream sales.

### Lurking Variables

Lurking variables are hidden variables that are correlated to the variables being studied. For example, a better explanation for the relationship between ice cream sales and drowning deaths is that during the summer, the hot temperature lead to increased ice cream sales and more people swimming. Time of the year would be considered a lurking variable.

### Misconceptions about categorical variables

Some common misconceptions about categorical variables are that there can be a correlation when one or both of the data set is comprised of categorical data. In this case, it does not make sense to find a correlation between the variables, rather an association is appropriate. For example, given data for the variables gender, and income, we cannot find a correlation between gender and income. But we can find an association in this case.



## Using a TI-83/84 Calculator to Find the Linear Regression Equation

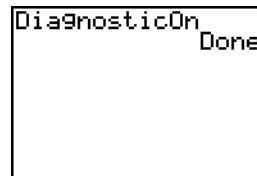
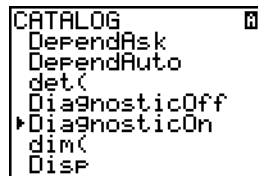
We will use the following example data set to introduce how to use a graphing calculator to find the best fit line equation and the correlation coefficient of a given data set.

Table 3.3 Education and Income Data

Years of Education	Income Per Year (in thousands)
19	110
20	125
16	60
16	70
18	90
12	15
14	35
12	25
16	50
17	70

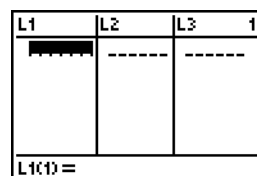
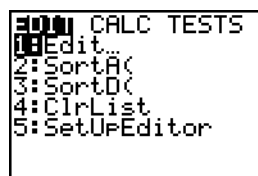
- Before entering our data set into the calculator, first turn on the diagnostics so our correlation coefficient  $r$  will be shown. Do this by pressing  $\boxed{2\text{nd}}$   $\boxed{[\text{CATALOG}]}$  then press  $\boxed{x^{-1}}$  then scrolling down to DiagnosticOn and pressing  $\boxed{\text{ENTER}}$  twice.

Figure 3.2 Calculator Diagnostic



- Enter the data set into the calculator. Do this by pressing  $\boxed{\text{STAT}}$  and press  $\boxed{\text{ENTER}}$  when Edit... is highlighted. This will take you to the lists screen. L1, L2, L3 are different lists of data, in our case we will put our Years of Education data in L1 and Income in L2.

Figure 3.3 Calculator Listing



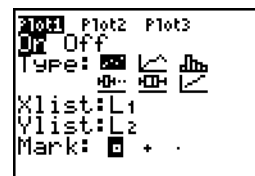
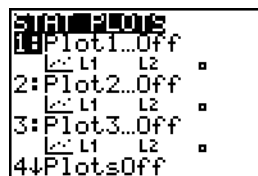
- To clear the data lists, press **[STAT]** and press **[ENTER]** when Edit... is highlighted. This will take you to the lists screen. To clear a list, use the up arrow to place the cursor on the list name then press **[CLEAR]** and **[ENTER]**. Take note that pressing **[DEL]** instead of **[CLEAR]** will delete the list from your calculator rather than clearing the data in the list.
- Enter the data Years of Education into L1 then right arrow to the next column and enter the Income data into L2. After, press **[2nd]** **[MODE]** to quit the listing screen.  
*Note: The lists must be the same length (L1 and L2 should have the same number of data in each). If they are not the same length then, ERR: DIM MISMATCH will be displayed when attempting to graph the data or performing the regression.*

Figure 3.4 Data Lists

L1	L2	L3	1
19	110		
20	125	-----	
16	60		
16	70		
18	90		
12	15		
14	35		
L1D=19			

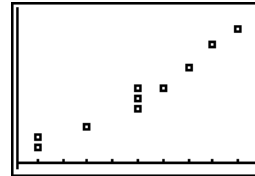
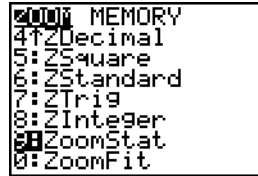
- To generate a scatterplot press **[2nd]** **[Y=]**, this will give you the STAT PLOTS menu. Turn on and define Plot 1 by highlighting 1 then pressing **[ENTER]**. This will bring you to the Plot1 screen. Highlight On and press **[ENTER]**. Again, press **[2nd]** **[MODE]** to quit the current screen. *Note: In defining Plot 1, ensure that the Xlist and Ylist are referring to the lists where your data are stored, in this case L1 and L2.*

Figure 3.5 Stat Plots



- Now, to view the scatterplot press **[ZOOM]** **[9]** or scroll down to 9: ZoomStat and press **[ENTER]**. Have the students observe the scatterplot and determine if there is a linear trend, and if there is a positive, negative or no correlation. Also have students guess the strength of the relationship by guessing the correlation coefficient value. When done, quit the screen.

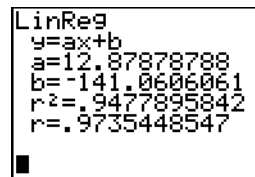
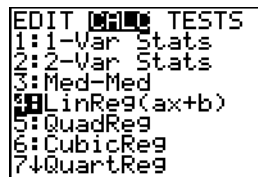
Figure 3.6 Zoom Stat



- To find the correlation coefficient, we will need to find the linear regression equation. To do this press **[STAT]** and scroll to the right to **CALC**. Press **[4]** **[ENTER]**. This will give the linear regression equation by giving the values a, b, as well as the correlation coefficient r. Have the students write their equation and interpret the correlation coefficient. Was there a strong linear relationship? *Note: The calculator defaults to L1=x, and L2=y. When using other lists, specify which lists are being used. LinReg xlist, ylist. In this case “xlist” should be the list where your x data is stored, and ylist is where your y data is stored. If you want to store the regression line for future use, this can be done by adding a Y function. For example, to store the regression line of L1 and L2 into function Y<sub>1</sub>, state LinReg L1, L2, Y<sub>1</sub>.*

Y<sub>1</sub> can be found by pressing **[VARS]** **[▶]** to Y-VARS, then **[ENTER]** **[ENTER]**.

Figure 3.7 Linear Regression



- As a reminder, when starting a new problem, it is a good idea to clear previous lists to minimize confusion and error. This can be done the same way we cleared data in step 3.

### Teaching Extension: Coefficient of Determination

The coefficient of determination is denoted by  $R^2$  or  $r^2$  as seen on the linear regression output of the TI-83/84 calculator. This checks the goodness of fit of a model, by determining how well the regression line approximates the actual data points.

The coefficient of determination is defined by the following formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \text{ where,}$$

$$\text{Residual sum of squares: } SS_{res} = \sum_i (y_i - f_i)^2$$

$$\text{Total sum of squares: } SS_{tot} = \sum_i (y_i - \bar{y})^2$$

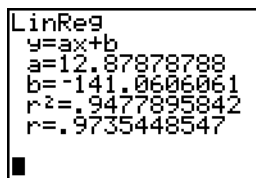
*Note:  $\bar{y}$  is the mean of the observed data.*

$R^2$  takes on the value between 0 and 1, representing the proportion of variance explained by the model. When  $R^2$  is 1 then the model fits perfectly, explaining all variability in  $Y$ . In general, the higher the  $R^2$ , the better the model fits the data. The higher the proportion of the explained variance, the closer the data points will fall to the linear regression model (or any model). Because of its definition in terms of the residual sum of squares,  $R^2$  can be used to assess goodness of fit for models other than linear models, where correlation is undefined.

For our learning tasks, we are working with simple linear regression, with cases of a single regressor. In this case,  $R^2$  is simply the square of the correlation coefficient  $R$ .

Figure 3.8 shows the linear regression calculator output from the class example. Notice that  $r^2$  is simply  $r$  that has been squared. In this example, we have  $r^2 = .9478$ , which can be interpreted as 94.78% of the variation in the income per year can be explained by the education level of people.

Figure 3.8 R-Squared Example



### 3.2 Student Worksheet: TV Watching Hours and Test Grades

Name \_\_\_\_\_ Date \_\_\_\_\_

#### Learning Task: TV Watching Hours and Test Grades

##### Common Core State Standards

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

**MCC9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**MCC9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient of a linear fit.

**MCC9-12.S.ID.9** Distinguish between correlation and causation.

##### Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them.**
2. **Reason abstractly and quantitatively.**
3. **Construct viable arguments and critique the reasoning of others.**
4. **Model with mathematics.**
5. **Use appropriate tools strategically.**

#### Part I:

1. Students in Ms. Garth's Algebra II class wanted to see if there are correlations between test scores and height and between test scores and time spent watching television. Before the students began collecting data, Ms. Garth asked them to predict what the data would reveal. Answer the following questions that Ms. Garth asked her class.
  - a. Do you think students' heights will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation?

*Answers may vary, but a possible answer could be: "I do not think there will be correlation between height and test grades, since it is not reasonable to think a person's height affects their intelligence or effort level."*

- b. Do you think the average number of hours students watch television per week will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation? Do watching TV and low test grades have a cause and effect relationship?

*Answers may vary, but a possible answer could be: “I think the average number of hours a student watches television will be negatively correlated with the student’s test grades. It is reasonable to think that the more TV you watch, the less time you spend studying, resulting in low test grades. However, it does not seem like these variables will be strongly correlated, since some people do not watch TV but do not spend time studying either. On the other hand, some students may watch a lot of TV and still study a lot.” Discuss correlation vs. causation with students. Give samples of variables that correlate and have them justify their argument.*

2. The students then created a table in which they recorded each student’s height, average number of hours per week spent watching television (measured over a four-week period), and scores on two tests. Use the actual data collected by the students in Ms. Garth’s class, as shown in the table below, to answer the following questions.

Table 3.4 Ms. Garth’s Class Data

Student	1	2	3	4	5	6	7	8	9	10	11	12	13
Height (inches)	60	65	51	76	66	72	59	58	70	67	65	71	58
TV hrs/week	30	12	30	20	10	20	15	12	15	11	16	20	19
Test 1	60	80	65	85	100	78	75	95	75	90	90	80	75
Test 2	70	85	75	85	100	88	85	90	90	90	95	85	85

- a. Which pairs of variables seem to have a positive correlation? Explain.

*Test 1 scores and test 2 scores appear to be positively correlated. For the most part, student performance on both tests was fairly consistent, so students who did well on test 1 also did well on test 2, while those who did not do well on test 1 didn’t do very well on test 2 either.*

- b. Which pairs of variables seem to have a negative correlation? Explain.

*Test 1 scores and hours per week watching television, and test 2 scores and hours per week watching television appear to be negatively correlated. In general, students who spent more time watching television had lower test scores than those who spent less time watching television.*

- c. Which pairs of variables seem to have no correlation? Explain.

*Height and hours per week watching television, test 1 scores and height, and test 2 scores and height seem to have no correlation. Height does not seem to be correlated with any of the other variables. That is, taller students do not seem to watch any more or less television or perform any better or worse on tests than shorter students.*

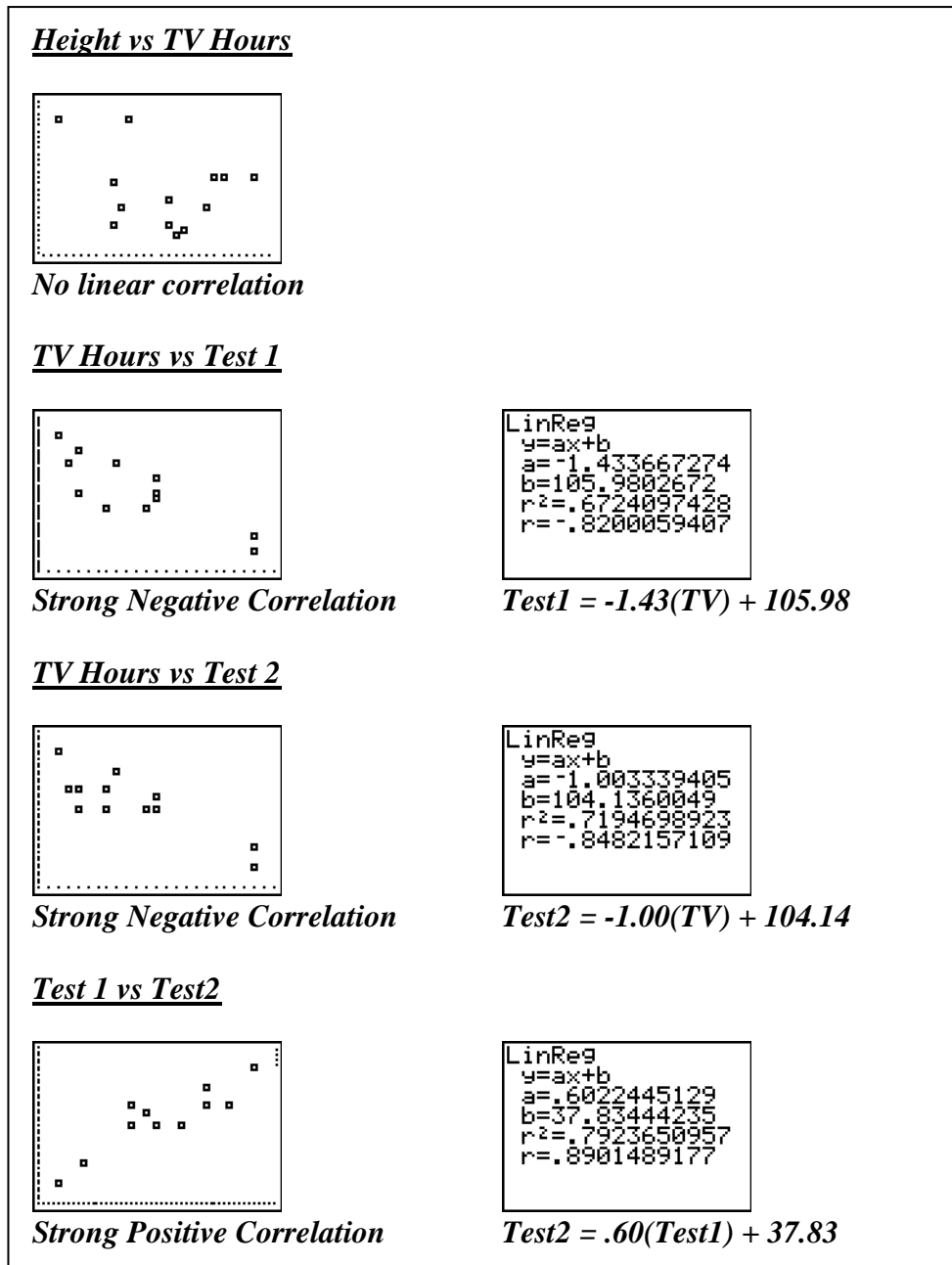
**Part II:**

9. For each pair of variables listed below, create a scatter plot with the first variable shown on the x-axis and the second variable on the y-axis. Are the two variables correlated positively, correlated negatively, or not correlated? What is their best-fit regression equation?

*Table 3.5 Correlation of Variables*

X	Y	Linear Correlation?	Linear Reg. Equation	Correlation coefficient
Height	TV Hrs			
TV Hrs	Test 1			
TV Hrs	Test 2			
Test 1	Test 2			

Figure 3.9 Correlation Calculator Outputs



10. From Ms. Garth's class data, and the correlations calculated, which variables can you say affects test scores?

*Students should be able to conclude that there is no linear relationship between height and test scores. But that the number of hours of watching TV has a strong negative correlation with test score.*



11. Maria has asked you to predict her Test 1 score and told you that she is 61 inches tall. What do you think her score will be and how much belief do you have in this?

***Answers will vary. The student should state that they do not believe in their prediction since there is no linear correlation between height and test score 1. Height is not a good indicator of test 1 score.***

12. On the other hand, Johnny says he does not know his height but he knows he watches 15 hours of TV per week. What do you think his score will be and how much belief do you have in this?

***Answers will vary. The student should state that they believe this is a good prediction of test score 1 since there is a strong negative correlation between TV hours watching and test score 1.***

13. Lauren made the conclusion that watching TV causes lower test scores. Can she make this conclusion? If not, why not?

***The student should state that watching TV does not cause lower test scores, but that there is a negative correlation between the number of hours of watching TV and the test score. Students who watched a higher number of hours of TV had lower test scores than those who watched less number of hours of TV.***

14. Jacob concluded that a student scoring well on test 1 result in a high score on test 2. Is this a valid conclusion?

***The student should state that scoring well on test 1 does not result on a high score for test 2, but that there is a positive correlation between test score 1 and test score 2, students who scored well on test 1 also scored well on test 2..***

15. What is your best-fit linear regression equation for TV Hours vs Test 2 scores? Interpret the slope of your equation in terms of the context.

***Best-fit linear regression equation:  $Test\ 2 = -1.00(TV) + 104.14$   
The slope of  $-1.00$  means that for every hour of TV watched, the test score decreases by 1 point.***

16. **Extension Question:** Interpret the coefficient of determination ( $r^2$ ) for the variable pairs, TV Hours vs Test 1, TV Hours vs Test 2, and Test 1 vs Test 2.

*Referring to the calculator outputs from Figure 3.8 we have the following coefficient of determination for the variable pairs.*

*TV Hours vs Test 1:  $r^2 = .6724$*

*TV Hours vs Test 2:  $r^2 = .7195$*

*Test 1 vs Test 2:  $r^2 = .7924$*

*62.74% of the variance in Test 1 scores is accounted for using the TV Hours watched.*

*71.95% of the variance in Test 2 scores is accounted for using the TV Hours watched.*

*79.24% of the variance in Test 2 scores is accounted for using Test 1 scores.*

### 3.3 Comparison Summary

A careful review of the original learning task showed standards indicated to be addressed by the task that was not satisfied. There were seven standards to be addressed – represent data on a scatter plot, fit a function to the data to solve problems in the context, emphasizing linear and exponential models, informally assess the fit of a function by plotting and analyzing residuals, fit a linear function for a scatter plot, interpret the slope and intercept of a linear model in the context of the data, compute and interpret the correlation coefficient of a linear fit, and distinguish between correlation and causation.

However, only three of these standards are covered in the original learning task. The students did not analyze residuals, interpret slope and intercept of their linear model, nor fit exponential models. In the original learning task students are to fit a function to the data and compute the correlation coefficient. These were the standards addressed in the original learning task and are still in the revised learning task. In addition to the previously addressed standards, the revised learning task also covers the distinction between correlation and causation, and interpretation of the slope of the linear model.

A third learning task is created to cover the CCGPS standard of fitting data into an exponential model, which was not previously addressed in the first two tasks.

## CHAPTER 4

### LEARNING TASK 3: U.S. POPULATION GROWTH

A third learning task was created to address the standard of fitting an exponential model to the data set. Exponential models were not introduced in any learning tasks but appeared as a performance task in the original tasks. This task was not a revision, rather an original task created to supplement the previous two tasks. The goal of this task is for students to be introduced to exponential models, and determine if a linear model or an exponential is a better fit for the data.

This task uses U.S. Population data, and asks the students to determine if they think a linear model or an exponential model best fits the data. It also shows that taking a partial data set may mean concluding the data is best represented as a linear model, while in actuality the best-fit model is an exponential model when considering the whole data set. The data table given is an example of data which citizens encounter in their life through newspaper articles. By having the students work with real-life data, students can see and understand the use of statistics in their everyday life. Students are asked to use their model to predict the population in a future date.

## 4.1 Teaching Guide

### Learning Task 3: U.S. Population Growth

#### Mathematical Goals

- Fit a linear and exponential model to the data and determine the appropriate model for the given data

Table 4.1 Learning Task 3 Standards

Common Core GPS	Student Worksheet Questions
<b>MCC9-12.S.ID.6</b> Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.	1
<b>MCC9-12.S.ID.6a</b> Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models.	2, 3
<b>MCC9-12.S.ID.6b</b> Informally assess the fit of a function by plotting and analyzing residuals.	6
<b>MCC9-12.S.ID.6c</b> Fit a linear function for a scatter plot that suggests a linear association. Interpret linear models.	2
<b>MCC9-12.S.ID.7</b> Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.	4

#### Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.

#### Introduction

In this third learning task, students will focus on exponential models. In doing this task, students analyze data sets, create scatter plots, determine the appropriate mathematical model, and justify their model selection.

This task provides a good example of how data points can appear to be linear over a relatively small domain, but how a different type of mathematical model might be more appropriate over a larger domain. This is an opportunity for students to discuss strengths and limitations of using mathematical functions to model real data. One discussion might arise as to whether other types of mathematical functions might sometimes be used for different types of data, perhaps leading students to look for patterns in data they might gather from sources like newspapers or books of world records.

**Time Required**

1 class period

**Materials**

Pencil and (graphing) paper; graphing calculator or statistical software package.

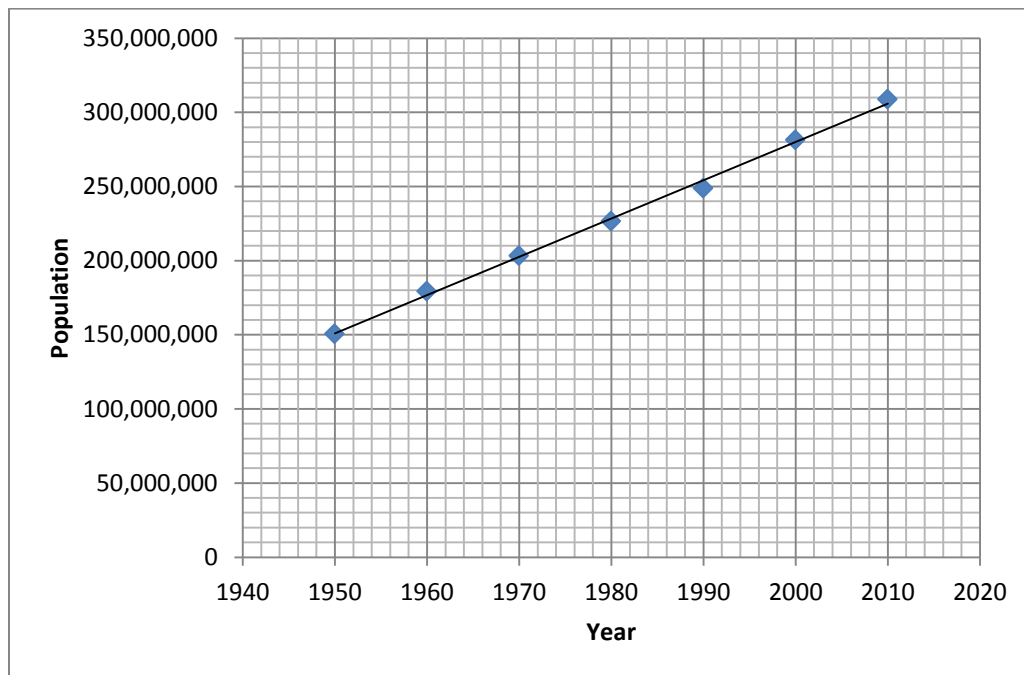
**Procedure:**

1. Recall the previous Learning Task – TV/Test Grades. Remind students how to use their graphing calculator for creating scatter plots and finding linear regression models.
2. Review different equation models and their graphs, such as the linear model, quadratic model, and exponential model. Show the students these graphical representations and make sure they are able to recognize the model given a graph.
3. Using the following part of the population data as an example, remind students how to create a scatter plot and fit a line to the data. Introduce exponential models, and how to use the calculator to fit an exponential model to the given data set.

Table 4.2 Partial Population Data

Year	Population
1950	150,697,361
1960	179,323,175
1970	203,302,031
1980	226,545,805
1990	248,709,873
2000	281,421,906
2010	308,745,538

Figure 4.1 Partial Population Data Scatter Plot



4. Have the students answer their U.S. Population Growth Learning Task worksheet.
5. Discuss with the students their conclusion regarding the linear and exponential model. Which do the students think is a better fit for the data? How does this compare to the example presented where only the data from 1950 to 2010 are considered? Students should understand that data points can appear to be linear over a relatively small domain; however a different type of mathematical model might be more appropriate over a larger domain.
6. **Extension:** After the task, ask the students to consider modeling the data using two regression models rather than one. Discuss with the students where they think they should split the data using some justification for their choice. For example, other variables influencing the data. Work with the students on finding two regression models for the data using the calculator. Find what is the more appropriate model – the linear model or exponential model by graphing residual plots.

## 4.2 Student Worksheet: U.S. Population Growth

Name \_\_\_\_\_ Date \_\_\_\_\_

### Learning Task: U.S. Population Growth

The data table shows the population in the United States gathered by the U.S. Census Bureau from 1800 to 2010. Use the data table to complete the task. Answer all questions in depth to show your understanding of the standards.

Table 4.3 U.S. Population Data

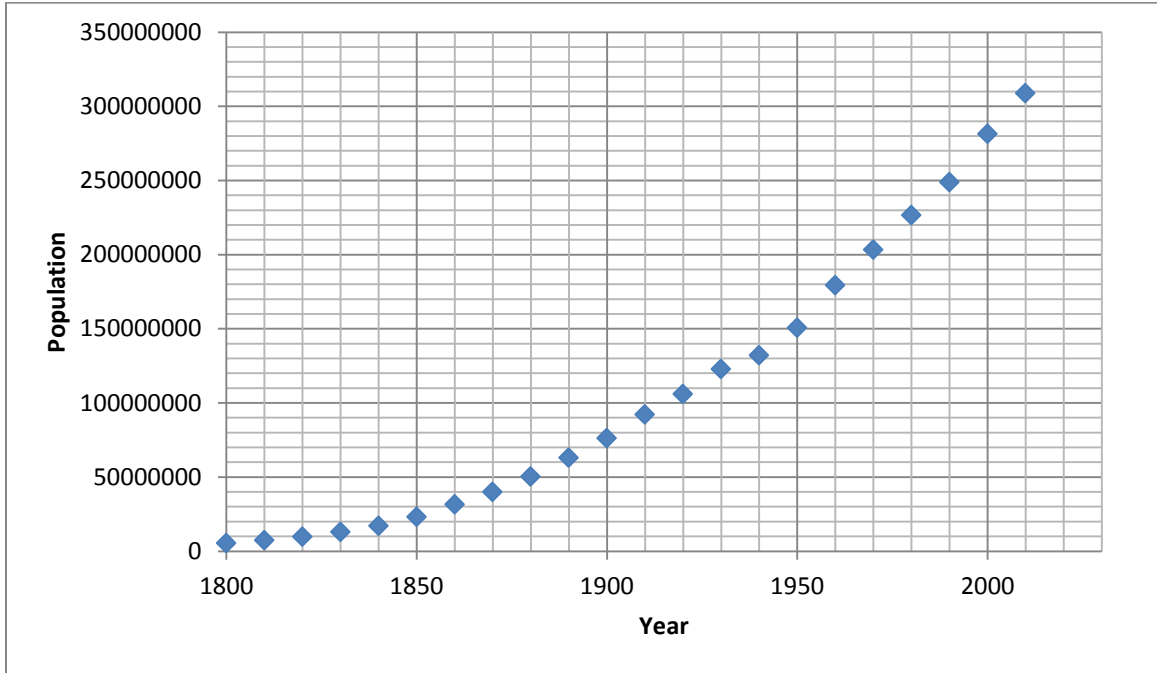
Year	Population
1800	5,308,483
1810	7,239,881
1820	9,638,453
1830	12,866,020
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	39,818,449
1880	50,189,209
1890	62,947,714
1900	76,212,168
1910	92,228,496
1920	106,021,537
1930	122,775,046
1940	132,164,569
1950	150,697,361
1960	179,323,175
1970	203,302,031
1980	226,545,805
1990	248,709,873
2000	281,421,906
2010	308,745,538

1. Create a scatter plot of the given data set of population over time.

*Use the same method of directions for students for graphing the data as given in the previous learning task (TV/Test Grades). The following graph from excel can be used for a clearer picture of the scatter plot to the class.*

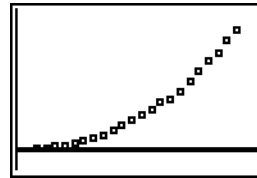


**Figure 4.2 Population Scatter Plot**



**Figure 4.3 Scatter Plot Calculator Output**

L1	L2	L3	1
1800	5.31E6	-----	
1810	7.24E6		
1820	9.64E6		
1830	1.29E7		
1840	1.71E7		
1850	2.32E7		
1860	3.14E7		
L1={1800, 1810, 1...			



- Fit a linear model to the data using your graphing calculator. Does this model seem to be the most appropriate model for this data set?

*Use the same method of directions for students for finding the linear regression model as given in the previous learning task (TV Watching Hours and Test Grades). Students' answers will vary about the appropriateness of the model. They should point out the curvature of the model.*

**Figure 4.4 Linear Regression Calculator Output**

```

LinReg
y=ax+b
a=1423906.919
b=-2604003573
r²=.9300285147
r=.9643798601
    
```

3. Stacy believes the data does not look linear, and thinks that an exponential model might be more appropriate for the data. Find the exponential model for the data. Do you think this is a more appropriate model?

*Use the ExpReg function on the graphing calculator to find the exponential model for this data. Students' answers will vary about the appropriateness of the model. They should point out the curvature of the model. (Note: The calculator really fits the exponential model as  $\ln y = \ln a + b x$ ; this is how a value for  $r$  becomes reasonable)*

**Figure 4.5 Exponential Regression Calculator Output**

```
ExpReg
y=a*b^x
a=1.1646887E-8
b=1.01918748
r^2=.9619268551
r=.9807786983
```

4. Compare your models with the model in the example given in class. Do the models agree? Why do you think there is a difference?

*Students should notice that the linear model was best during the example, but in this task the exponential model turned out to be more appropriate. Point out to students that sometimes having a small data set from a small domain is not enough data to create a model for analysis and drawing conclusions from it is not appropriate.*

5. Use both models to predict the population in year 2020. Which model makes sense in representing the growth of the US Population?

**Linear: Population = 1423906.92 \* year – 2604003573**

**Exponential: Population = .000000011646887 \* 1.01918748<sup>year</sup>**

*At year 2020, the US Population is as follows:*

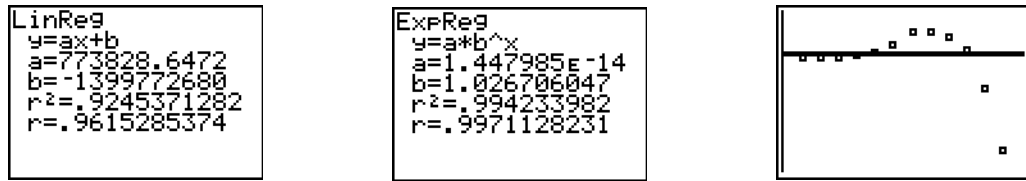
**Linear: Population = 272, 288, 403.4**

**Exponential: Population = 548,940,715.2**

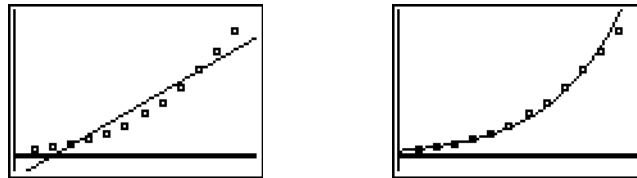
*Given that the US Population is 308,745,538 from the data table provided, and considering the increasing trend in population, then the linear model doesn't make sense in this case, giving us the exponential model as the better model for representation of the data provided.*

6. **Extension:** Fit two models to the population data. Changes in immigration laws occurred after World War 1 in 1918, so consider 1910 and 1920 as the splitting points of the data. Fit a linear and exponential model to both of the data subsets (1800-1910, and 1920-2010) and decide which is model is more appropriate by graphing the residual plots.

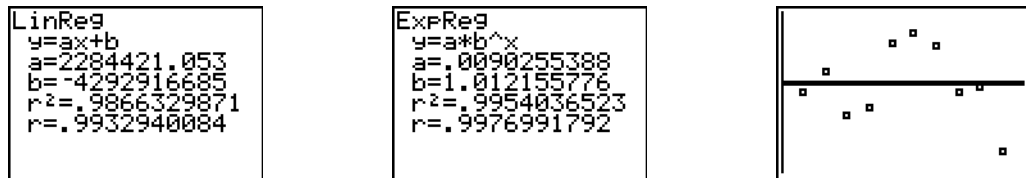
**Figure 4.6 Pre-War Regression Models and Residual Plot**



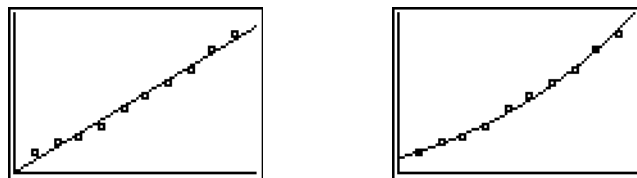
**Figure 4.7 Pre-War Scatter Plot with Model Fits**



**Figure 4.8 Post-War Regression Models and Residual Plot**



**Figure 4.9 Post War Scatter Plot with Model Fits**



*Both residual plots show some pattern, suggesting that a linear regression model isn't appropriate for the data. This can be confirmed from the higher value of the correlation coefficient  $r$  for the exponential models, and visually from the scatter plots with the exponential model fitted.*

### 4.3 Summary

This task was created to place emphasis on exponential models. The previous two tasks only involved linear models. CCGPS standards states “Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models”.

This third task asks the students to fit both a linear and exponential models to a given data set and determine the appropriate model for the given data. The students are asked to compare the models fitted for the whole data set versus a small subset of the data. This provides the student the emphasis of linear and exponential models as given in the standards. The students are asked to predict the population in a future year and asked to make sense of the prediction based on the model they chose.

An extension to this task is given. The students are to fit two regression models to the data, and using residual plots to determine if a linear or an exponential model is more appropriate.

The three tasks cover all of the CCGPS standards pertaining to representation and interpretation of data on quantitative variables and linear models. A performance task is given as a culmination of the topics introduced in the three learning tasks. This will assess the students understanding of the content taught in the previous three tasks.

## CHAPTER 5

### PERFORMANCE TASK: SAVINGS ACCOUNTS

The Savings Accounts performance task assesses students' understanding of regression models by determining the best-fit model and interpreting the model in the context given.

The original performance task given in Appendix C, Equal Salaries for Equal Work asks the students to create a scatter plot of the data given, fit a linear and exponential model, then determine which model better fits the data. The revised task asks the students to do the same, however using a data set consisting of savings accounts balances for three different savings plans. This data set is another way students can see the usefulness of statistics in their everyday life, as they compare different saving plans for the best investment.

In the revised version of the task, the students are asked to interpret their model in the context given. For example, compare the following questions:

Original question: “Terry and Tomas are trying to decide what type of model will most accurately represent the data. Terry thinks that a linear model might be most appropriate for the scatter plot. Help Terry find reasonable function rules for each scatter plot. Explain how you found these.”

Revised question: “Terry and Tomas are trying to decide what type of model will most accurately represent the data. Terry thinks that a linear model might be most appropriate for each scatter plot. Help Terry find the linear regression model for each of the data sets, and interpret the slopes of the models with respect to the context.”

The new question not only tests the students’ ability to find the best-fit linear model for the data set, but also their understanding of what the model represents by asking the students to interpret the slope of the models they find.

The changing of the scenario and adding question eight on the student worksheet lets students do some interesting analysis given a data set. This question solicits students understanding of how they can use given data presented to them to make citizenship decisions pertaining to their lives. It tests the students' ability to interpret data and make informed decisions from it.

## 5.1 Teaching Guide

### Savings Accounts

#### Mathematical Goals

- Fit a linear and exponential model to the data and determine the appropriate model for the given data

Table 5.1 Performance Task Standards

Common Core GPS	Student Worksheet Questions
<b>MCC9-12.S.ID.6</b> Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.	1
<b>MCC9-12.S.ID.6a</b> Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models.	2, 4, 6
<b>MCC9-12.S.ID.6b</b> Informally assess the fit of a function by plotting and analyzing residuals.	9
<b>MCC9-12.S.ID.6c</b> Fit a linear function for a scatter plot that suggests a linear association. Interpret linear models.	2, 3
<b>MCC9-12.S.ID.7</b> Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.	2
<b>MCC9-12.S.ID.8</b> Compute (using technology) and interpret the correlation coefficient of a linear fit.	7

#### Common Core State Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them.**
2. **Reason abstractly and quantitatively.**
3. **Construct viable arguments and critique the reasoning of others.**
4. **Model with mathematics.**
5. **Use appropriate tools strategically.**
6. **Attend to precision.**
7. **Look for and make use of structure.**
8. **Look for and express regularity in repeated reasoning.**

#### Introduction

This performance task asks students to compare additive and multiplicative growth (represented by linear and exponential models) to make predictions and solve problems within the context of gender-based salary differences. In doing this task, students analyze data sets, create scatter plots, determine the most appropriate mathematical model, and justify their model selection.

This task provides a good example of how data points can appear to be linear over a relatively small domain, but how a different type of mathematical model might be more appropriate over a larger domain. This is an opportunity for students to discuss strengths and limitations of using mathematical functions to model real data. One discussion might arise as to whether other types of mathematical functions might sometimes be used for different types of data, perhaps leading students to look for patterns in data they might gather from sources like newspapers or books of world records.

**Prerequisites**

Students must have knowledge of using the graphing calculator to create linear and exponential models and to analyze residuals. It is important that students understand how to assess the fit of a function to data and choose a function suggested by context.

**Learning Targets**

When making statistical models, technology is valuable for varying assumptions, exploring consequences and comparing predictions with data. Students will interpret the correlation coefficient and show understanding of strengths and limitations of using mathematical functions to model real data.

**Time Required**

1 class period

**Materials**

Pencil and (graphing) paper; graphing calculator or statistical software package.

## 5.2 Student Performance Task Worksheet

Name \_\_\_\_\_ Date \_\_\_\_\_

### Performance Task: Teacher Salary

Use the data table to complete the task. Answer all questions in depth to show your understanding of the standards.

Table 5.2 Savings Chart

<b>Year</b>	<b>John</b>	<b>Lucy</b>	<b>Mark</b>
0	1,000	10,000	5,000
1	3,628	12,826	10,274
2	6,288	15,709	15,601
3	8,982	18,650	20,982
4	11,710	21,650	26,417
5	14,472	24,711	31,906
6	17,269	27,834	37,450
7	20,101	31,020	43,050
8	22,968	34,270	48,707
9	25,782	37,585	54,420
10	28,813	40,968	60,191
11	31,790	44,419	66,019
12	34,804	47,939	71,906
13	37,857	51,531	77,853
14	40,948	55,195	83,859
15	44,078	58,933	89,925
16	47,078	62,747	96,052
17	50,456	66,637	102,241
18	53,705	70,606	108,492
19	56,995	74,655	114,806
20	60,326	78,786	121,183
21	63,700	83,000	127,624
22	67,116	87,299	134,130
23	70,574	91,685	140,702
24	74,076	96,160	147,339
25	77,623	100,725	154,043

Table 5.3 Savings Accounts

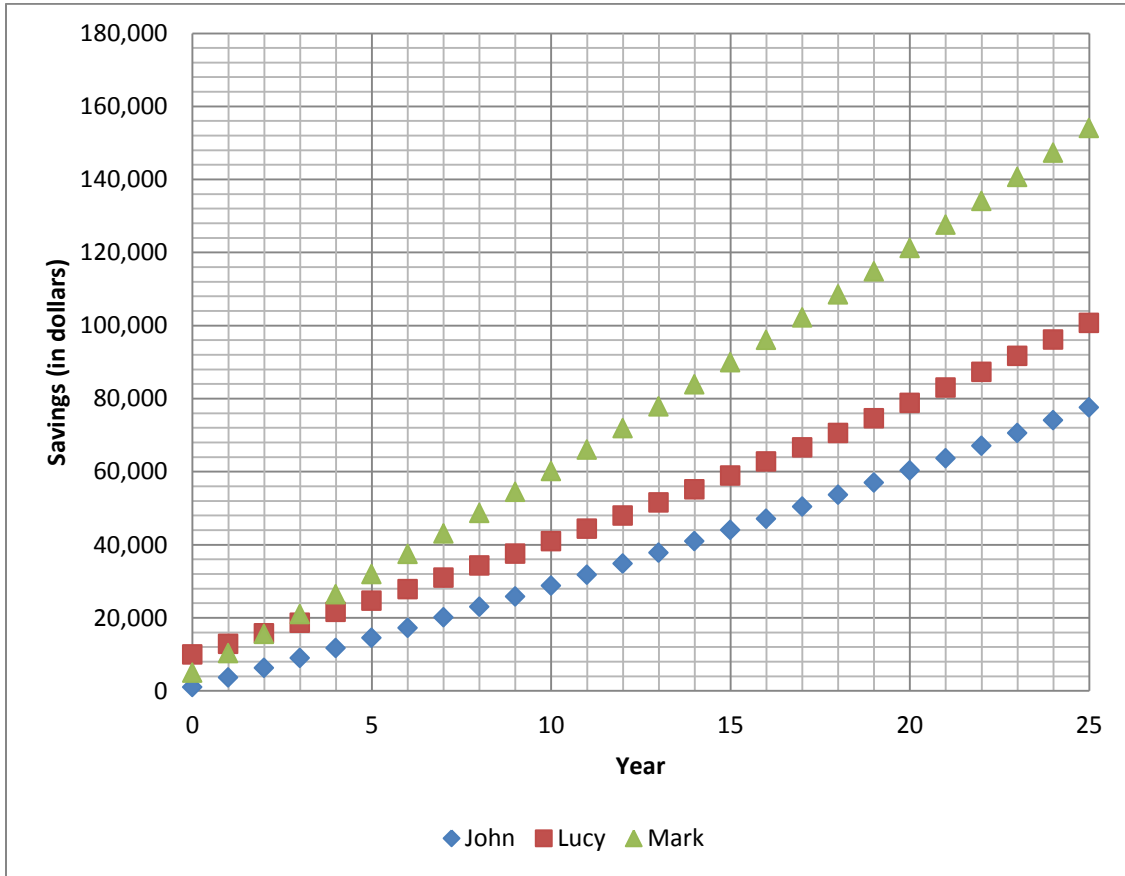
	<b>John</b>	<b>Lucy</b>	<b>Mark</b>
Initial Deposit (Principal)	1,000	10,000	5,000
Annual Interest Rate	1.25%	2%	1%
Contribution per pay period	100	100	200
Pay periods per year	26	26	26



1. Create three scatter plots of the data of savings amount throughout over age. Describe two things you notice about the scatter plots.

*This can be done on graphing paper or with the use of a graphing calculator. Use the calculator guide in the previous learning task – Simple Linear Regression for directions on how to create a scatter plot on the TI-84 graphing calculator.*

**Figure 5.1 Savings Accounts Scatter Plot**



2. Terry and Tomas are trying to decide what type of model will most accurately represent the data. Terry thinks that a linear model might be most appropriate for each scatter plot. Help Terry find the linear regression equation for each of the data sets, and interpret the slopes of the equations with respect to the context.

*Using a graphing calculator to determine a regression line, we have the following:*

**John:**  $y = 3060.33x - 864.18$

**Lucy:**  $y = 3615.73x + 6554.91$

**Mark:**  $y = 5956.25x + 2091.94$

*John's savings increase at a rate of \$3,060.33 every year, Lucy's savings increase at a rate of \$3,615.73 each year, and Mark's savings increase at a rate of \$5,956.25 each year. This shows that Mark's savings is increasing at a much faster rate than John's and Lucy's, with almost double of their change in savings value each year.*

- Using the linear models, will John's savings ever equal Lucy's savings? If so, at what year in their savings will this occur?

*Using the linear models created from the data provided, John's linear model has a smaller y-intercept and a smaller slope, while Lucy's linear model has a larger y-intercept and a larger slope. So Lucy's savings will increase at a faster rate, so John will not equal to Lucy's.*

- Tomas thinks that an exponential model might be most appropriate for each scatter plot. Help Tomas find the exponential function for each scatter plot.

*Using a graphing calculator to determine a regression line, we have the following:*

**John:**  $y = 5774.55 * (1.1302^x)$

**Lucy:**  $y = 15237.43 * (1.0872^x)$

**Mark:**  $y = 15289.77 * (1.1131^x)$

- Using the exponential models, will John's savings ever equal Lucy's savings? If so, when will this happen? Explain how you found your answer.

*Using the exponential models, John's savings will eventually equal Lucy's. The exponential model of John's savings has a base of 1.13012, and the exponential model of Lucy's savings has a base of 1.1178. Since John's model has a higher base, his earnings are increasing at a faster rate and will eventually surpass Lucy's savings. These functions can be graphed to determine their intersection. This occurs at  $x = 25.0146$ . This means that it will take about 25 years for John and Lucy to have the same amount in their savings account with a value of \$123,361.72.*

6. Based on the scatter plot and models, which of the linear model or exponential model do you think best represents the data? Why do you think so?

*The gaps between the scatter plots are not all equal, indicating a non-linear model. The compounded interest rate indicates an exponential model.*

7. Using the correlation coefficient, which of the savings accounts have the strongest correlation? Is this correlation positive, negative, or no correlation?

**Figure 5.2 John's Linear Model**

```
LinReg
y=ax+b
a=3060.331282
b=-864.1794872
r2=.9982216088
r=.9991104087
```

*The correlation coefficient for John's savings account is .9991, which means there is a strong positive correlation between time and the amount of money in her savings account.*

**Figure 5.3 Lucy's Linear Model**

```
LinReg
y=ax+b
a=3615.729915
b=6554.91453
r2=.9955561276
r=.9977755898
```

*The correlation coefficient for Lucy's savings account is .9978, which means there is a strong positive correlation between time and the amount of money in her savings account.*

**Figure 5.4 Mark's Linear Model**

```
LinReg
y=ax+b
a=5956.25094
b=2091.940171
r2=.9988828306
r=.9994412592
```

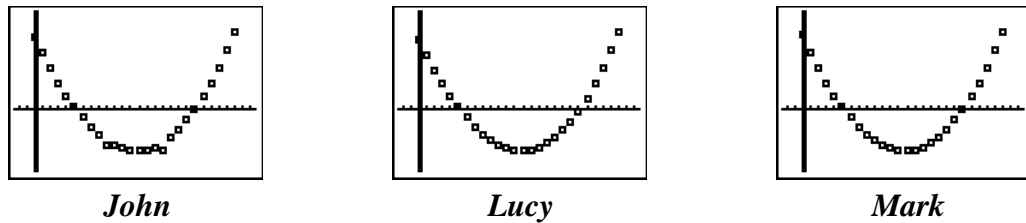
*The correlation coefficient for Mark's savings account is .9994, which means there is a strong positive correlation between time and the amount of money in his savings account.*

8. Using the data provided, consider the following scenario. Lucy decided to make an initial deposit of \$10,000 since she is starting at age 40, while John started his savings account at the age of 30. Also, Lucy decided to make the same contribution per pay period of \$100, but was able to get a higher annual interest rate on her savings account. She believes she would have more money than John when they retire at the age of 55. Is this belief true? Considering how long it took for John and Lucy to have equal savings, which savings plan do you think is better for saving more money?

*After 25 years of savings for John has \$77,623, while Lucy has \$58,933 in her savings after 15 years since she started her account at the age of 40. Lucy's account is better for savings due to its higher interest rate, however it is only better if she started at the same time or earlier than John.*

9. **Extension:** Using residual plots check the appropriateness of the linear regression model for each of the three savings accounts.

*Figure 5.5 Residual Plots*



*All three residual plots suggest that the linear regression model is not appropriate for modeling the different savings accounts. The curve in the residual plot shows a non-random pattern and indicates a higher order model is recommended, in our case this indicates the exponential model is a better fit to the data set.*

### 5.3 Comparison Summary

Considering the original performance task provided; the main problem of the use of this assessment as the introduction of exponential models. This issue has been addressed through the creation of the third learning task.

Another shortcoming of the original performance task was the lack of addressing the standards of plotting and analyzing residuals. This was considered in the revision of the assessment. However, in considering the time constraint for a class period assessment and the problem presented, it is best to remove these from the standards statement for this task.

The original performance task data of women's earnings versus men's earnings were replaced with data on different savings accounts. This data provides the student data that is indeed exponential, as was the intention of the original task, and solicits contextual understanding of the concepts presented.

Plotting and analyzing residuals as given in learning task 1 will take the students too much time to do. Also, students are not expected to plot data points of a large data set by hand as it is more efficient for students to know how to create a scatter plot using technology.

An extension question is given for students to check the appropriateness of their linear regression model using residual plots. Residual plots were introduced as an extension in the first learning task.

## CHAPTER 6

### CONCLUSION

In general, the revised tasks ensure the students gain statistical literacy by giving the students real-life data that is relevant to their lives. An emphasis is placed on relating the findings back to the context, and interpreting their findings such as the slope and intercept where appropriate of the best-fit line in terms of the data given. This allows the students to analyze the data in such a way that they can make conclusions and informed decisions regarding the data sets given. For example, by asking the students “What can you conclude about the relationship between the number of hours of TV students watch per week and your test score?” We ensure that the students not only meet the standard of fitting a linear function to the scatter plot and suggesting a linear association, but that they are able to do this in the context of the data given which shows the students quantitative literacy.

The revised tasks also provide the teachers with more content knowledge to guide them in teaching the lessons. As research has shown, many teachers do not have the statistical background knowledge to teach units such as regression. The guidance provided for the teachers is essential to the professional development of teachers, better informed teachers being able to provide a deeper understanding of the material for the students.

## REFERENCES

- Achieve (2013). Achieving the Common Core. Retrieved from <http://www.achieve.org/achieving-common-core>
- Achieve (2012). Achieve & The American Diploma Project Network. Retrieved from <http://www.achieve.org/files/About%20AchieveADP-Apr2012.pdf>
- Common Core State Standards Initiative (CCSSI). (2012a). Frequently Asked Questions. Retrieved from <http://www.corestandards.org/resources/frequently-asked-questions>
- Common Core State Standards Initiative (CCSSI). (2012b). In The States. Retrieved from <http://www.corestandards.org/in-the-states>
- Franklin, C. (2012) Chris Franklin Statistics Summer 2012 Videos: Introduction [Video Webcast] Retrieved from: <http://real.doe.k12.ga.us/vod/gso/math/CCGPS-Statistics/CCGPS-Statistics-Training-1.mp4>
- Georgia Department of Education (GA DOE). (2011). Frequently Asked Questions. Retrieved from <http://www.georgiastandards.org/Standards/Pages/BrowseStandards/BrowseGPS.aspx>
- Georgia State Board. (2010). Achieving the Common Core. Georgia State Board Report: Mathematics Findings. Retrieved from <http://eboard.eboardsolutions.com/meetings/Attachment.aspx?S=1262&AID=244381>
- Humphrey, P. Personal Communication, June 21, 2013
- Madison, B.L. (2003). Articulation and quantitative literacy: A view from inside mathematics. In B.L. Madison & L. A. Steen (Eds.) *Quantitative literacy. Why numeracy matters for schools and colleges* (pp. 153-164). Princeton, NJ: The National Council of Education and the Disciplines.
- Moreno, J.L. (2002). "Toward a Statistical Literate Citizenry: What Statistics Everyone Should Know," in *Proceedings of the 6<sup>th</sup> International Congress on Teaching Statistics*, ed. B. Phillips, July 7 – 12, 2002, Cape Town, South Africa, Voorburg, the Netherlands: International Statistical Institute.
- National Commission on Mathematics and Science Teaching for the 21<sup>st</sup> Century (National Commission). (2000). *Before It's Too Late*. Washington, DC: U.S. Department of Education.
- National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*, Reston, VA: Author
- No Child Left Behind (NCLB) Act of 2001, Pub. L. No 107-110. § 115, Stat. 1425 (2002).

- Scheaffer, R. L. (2003). Statistics and quantitative literacy. In B.L. Madison & L. A. Steen (Eds.) *Quantitative literacy. Why numeracy matters for schools and colleges* (pp. 145-152). Princeton, NJ: The National Council of Education and the Disciplines.
- Scheaffer, R. L. (1990). The ASA-NCTM Quantitative Literacy Project: An Overview. In D. VereJones (Ed.) *Proceedings of the Third International Congress on Teaching Statistics*. Dunedin, New Zealand: International Statistical Institute .
- Steen, Lynn Arthur. (2002). "Achieving Mathematical Proficiency for All." *College Board Review*. (Spring) 196:4-11
- Watkins, A. AP Statistics Listserve Email Communication, September 17, 2011



## APPENDIX A

### LEARNING TASK 1: SPAGHETTI REGRESSION

#### Spaghetti Regression

Adapted from: <http://txcc.sedl.org/events/previous/092806/10ApplyingStrategies/math-teks-alg1.pdf>

#### Mathematical Goals

- To investigate the concept of goodness of fit and develop an understanding of residuals in determining a line of best-fit

#### Common Core State Standards

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

#### Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.

#### Introduction

Students will investigate the concept of the “goodness-of-fit” and its significance in determining the regression line or best-fit line for the data. This is the first exploration in a series of three activities to explore a best-fit line and residuals. Fitting the graph of an equation to a data set is covered in all mathematics courses from Algebra to Calculus and beyond. The objective of this activity is to explore the concept in-depth.

In real life, functions arise from data gathered through observations or experiments. This data rarely falls neatly into a straight line or along a curve. There is variability in real data, and it is up to the student to find the function that best 'fits' the data. Regression, in its many facets, is probably the most widely used statistical methodology in existence. It is the basis of almost all modeling.

Students create scatter plots to develop an understanding of the relationships of bivariate data; this includes studying correlations and creating models from which they

will predict and make critical judgments. As always, it is beneficial for students to generate their own data. This gives them ownership of the data and gives them insight into the process of collecting reliable data. Teachers should naturally encourage the students to discuss important concepts such as goodness-of fit. Using the graphing calculator facilitates this understanding. Students will be curious about how the linear functions are created, and this activity should help students develop this understanding.

**Materials:**

- Spaghetti or linguine (3 or 5 pieces of spaghetti per student)
- Transparent tape (roll for each group)
- Transparencies of Overhead 1 and Measuring Notes
- Handouts – copy for each student of the Scatter plot,
- Student Activity: Spaghetti Regression, and Measuring Notes
- Rulers (optional)

**Grouping: 4-5 students per group**

**Time: 50 to 60 minutes**

<i>Procedures</i>	<i>Notes</i>
<p><b>1. Activity 1</b>  <i>Introduce the topic of goodness of fit with Overhead 1.</i></p> <p><i>Ask: Why do we say that the line in the top graph fits the points better than the line in the bottom graph?</i></p> <p><i>Can we say that some other line might fit them better still?</i></p> <p><i>Say: Usually we think of a close fit as a good fit. But, what do we mean by close?</i></p>	<p><i>Discuss the importance of modeling and lead student discussions of concepts such as goodness-of-fit, (See the Background information provided in this lesson.)</i></p>
<p><b>2. Give each student 3-5 pieces of spaghetti, the Scatter plot handout, and Student Activity: Spaghetti Regression.</b></p>	
<p><b>3. Have the students examine the plot and visually determine a line of best-fit (or trend line) using a piece of spaghetti. They</b></p>	<p><i>This should be done individually so that there is variation in the choice of lines within each group.</i></p>

<p><i>then tape the spaghetti line onto their graph as described in #1 on the Student Activity handout.</i></p>	
<p><b>4.</b> <i>Before students go on to #2 on the Student Activity handout, ask:</i></p> <p><i>Who has the best line in your group?</i></p> <p><i>How can we determine this? (Do not discuss how to measure this yet; this will be addressed later.)</i></p>	<p><i>This is the central idea behind linear regression. To determine a line-of best fit you must have an agreed upon measure of “goodness”. If that measure is “closeness of the points to the line”, the best line is then the line with the least total distance from the points to the line. There are many types of regression. The most common is the method of least squares.</i></p> <p><i>Intuitively, we think of a close fit as a good fit. We look for a line with little space between the line and the points it's supposed to fit. We would say that the best fitting line is the one that has the least space between itself and the data points which represent actual measurements.</i></p>
<p><b>5.</b> <i>Have the students follow the directions for #2 by using a second piece of spaghetti to measure the distance from each point to the line. Then break off that length.</i></p> <p><i>Groups may measure vertically, horizontally, perpendicularly, etc. However, each member of a group must measure the same way. It is very important for each group to decide their method for measuring before they begin.</i></p>	<p><i>Encourage diversity in measuring methods among the groups to add depth to the following discussions.</i></p>
<p><b>6.</b> <i>Have the students line up their spaghetti distances to determine who in their group has the closest fit. Then, they replace the segments and tape them to their scatter plot.</i></p>	<p><i>This will determine the total error (i.e., total distance from their line to the data). The scatter plot is on centimeter paper. To be able to express the total error as a numerical value you may want students to use a ruler.</i></p>
<p><b>7.</b> <i>Have each group present their</i></p>	<p><i>Discuss the fact that since the groups</i></p>

<p><i>method and results. A good way to accomplish this is to have the “winner” from each group come up to the front to do the reporting. They can then be grouped by their method of measurement. Have reporter share, discuss, compare, and contrast their results.</i></p>	<p><i>used different methods of measuring, they cannot determine best-of-fit for the entire class.</i></p> <p><i>Discuss accuracy of measurement.</i></p> <p><i>Did they measure from the edge of each point or the middle?</i></p>
<p><b>8.</b> <i>Hand out Measuring Notes and use it to discuss three ways (vertical, horizontal, and perpendicular) to measure the space between a point and a line.</i></p> <p><i>Discuss the meaning of a residual and why it is used in evaluating the accuracy of a model. Use the overheads of this page to cultivate the discussion.</i></p>	<p><i>Why measure vertically?</i></p> <p><i>The sole purpose in making a regression line is to use it to predict the output for a given input. The vertical distances (residuals) represent how far off the predictions are from the data we actually measured.</i></p>

## Group Learning Task: Spaghetti Regression

### Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.

### Common Core State Standards

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

1. Examine the plot provided and visually determine a line of best-fit (or trend line) using a piece of spaghetti. Tape your spaghetti line onto your graph.
  
2. Now investigate the “goodness” of the fit. Use a second piece of spaghetti to measure the distance from the first point to the line. Break off this piece to represent that distance. Each person at the table must measure in the same way, so discuss the method you will use before starting. Repeat this for each point in the scatter plot.

***Teacher notes: Encourage at least one group to use the shortest distance from the point to the line (i.e., the perpendicular distance.)***

3. Line up your “spaghetti distances” to determine who in your group has the “closest” fit. Determine the total error. (i.e., total distance from your line to the data.) Then replace the segments and tape them to your scatter plot.

Total error = \_\_\_\_\_ cm (nearest tenth)

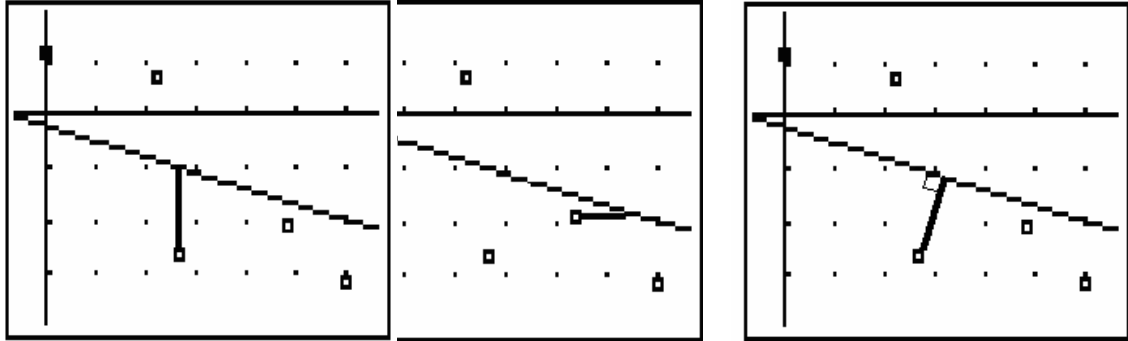
*Have each group present their method and results. A good way to accomplish this is to have the “winner” from each table come up to the front. They can then be grouped by their method of measurement. Have each share, discuss, compare, and contrast.*

*Discuss the fact that since the groups used different methods of measuring, we cannot determine best-of-fit for the entire class. Discuss the accuracy of their measurements. Did they measure from the edge of each point or the middle, etc.?*

*Use the page titled “Measuring Notes” to discuss three ways to measure the space between a point and the line. Discuss the meaning of a residual and why it is used in evaluating the accuracy of a model. Use the overheads of this page to cultivate the discussion.*

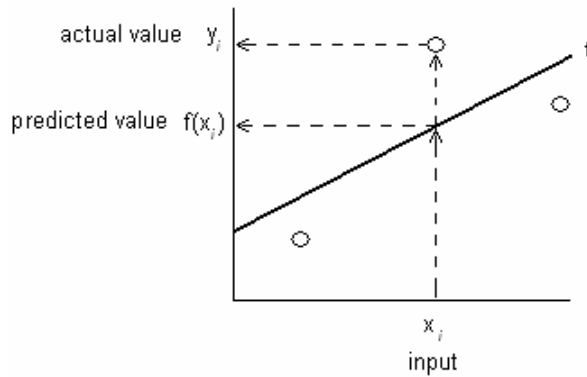
## Measuring Notes

There are at least three ways to measure the space between a point and the line: vertically in the y direction, horizontally in the x direction, and the shortest distance from a point to the line (on a perpendicular to the line.)



In regression, we usually choose to **measure the space vertically**. These distances are known as **residuals**.

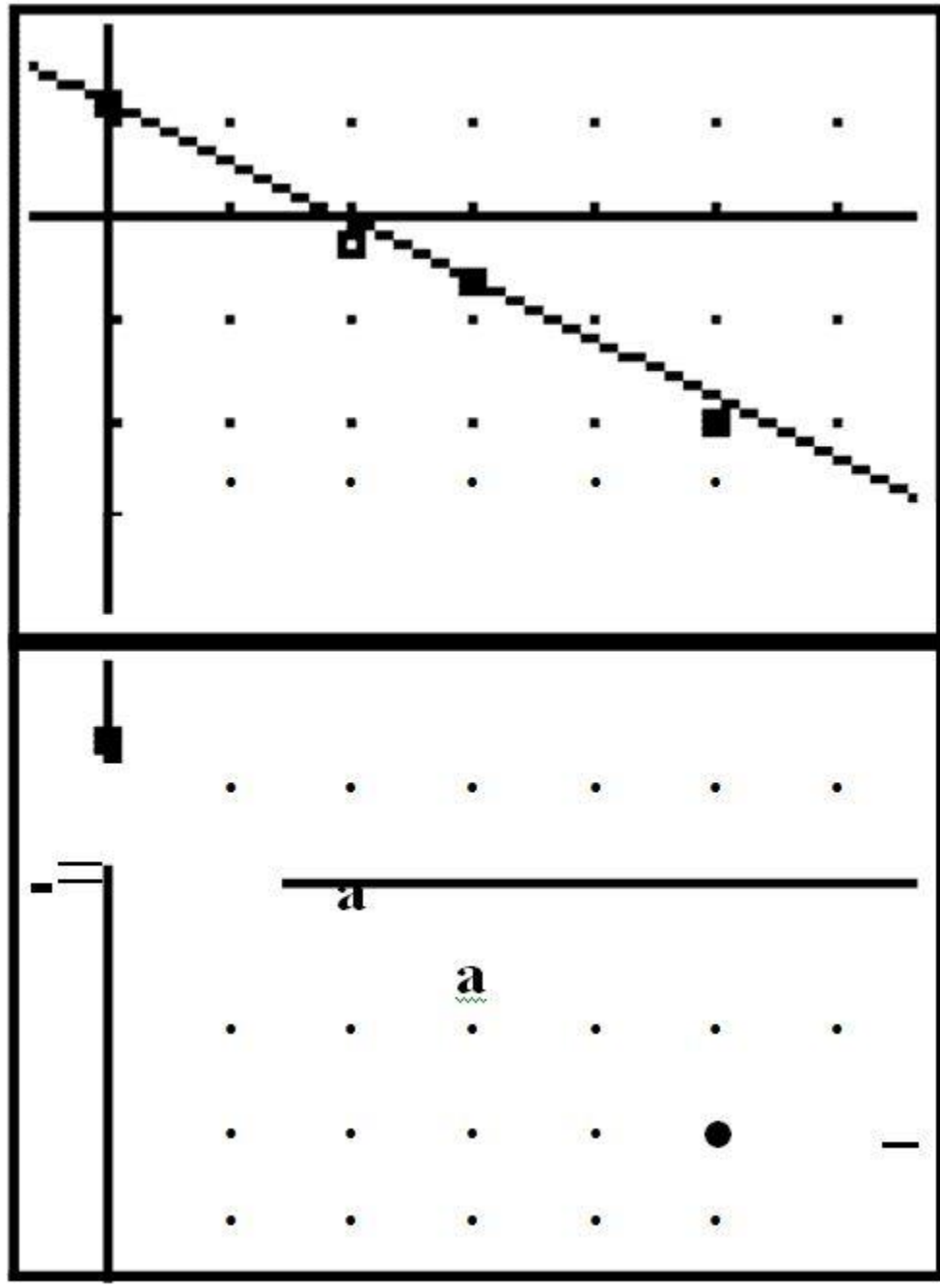
- Why would you want to measure this way? What do the residuals represent in relation to our function? Consider the purpose of the line and the following diagram.



The purpose of regression is to find a function that can model a data set. The function is then used to *predict* the y values (outputs or  $f(x)$ ) for any given input x. So, the vertical distance represents how far off the prediction is from the actual data point (i.e., the “error” in each prediction.) Residuals are calculated by subtracting the model’s predicted values,  $f(x_i)$ , from the observed values,  $y_i$ .

$$\text{Residual} = y_i - f(x_i)$$

# Overhead 1





## Group Learning Task: Spaghetti Regression

### **Common Core State Standards for Mathematical Practice**

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.

### **Common Core State Standards**

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

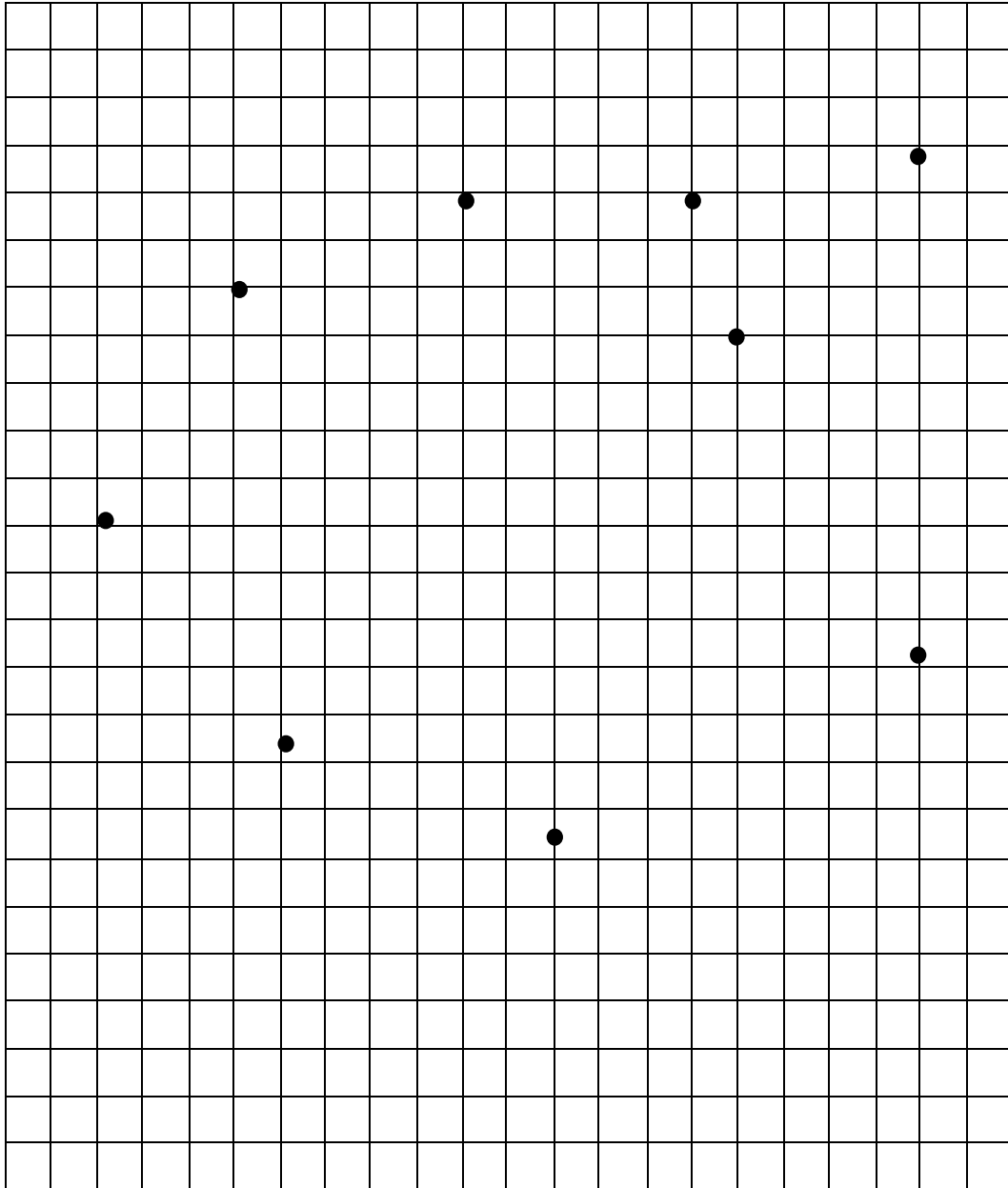
**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

1. Examine the plot provided and visually determine a line of best-fit (or trend line) using a piece of spaghetti. Tape your spaghetti line onto your graph.
  
2. Now investigate the “goodness” of the fit. Use a second piece of spaghetti to measure the distance from the first point to the line. Break off this piece to represent that distance. Each person at the table must measure in the same way, so discuss the method you will use before starting. Repeat this for each point in the scatter plot.
  
3. Line up your “spaghetti distances” to determine who in your group has the “closest” fit. Determine the total error. (i.e., total distance from your line to the data.) Then replace the segments and tape them to your scatter plot.

Total error = \_\_\_\_\_cm (nearest tenth)

# Scatter plot



## APPENDIX B

### LEARNING TASK 2: TV/TEST GRADES

#### TV/Test Grades

##### Mathematical Goals

- Represent data on a scatter plot
- Describe how two variables are related
- Informally assess the fit of a function by plotting and analyzing residuals
- Fit a linear function for a scatter plot that suggests a linear association

##### Common Core State Standards

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6a** Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

**MCC9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**MCC9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient of a linear fit.

**MCC9-12.S.ID.9** Distinguish between correlation and causation.

##### Standards for Mathematical Practice

1. **Make sense of problems and persevere in solving them.**
2. **Reason abstractly and quantitatively.**
3. **Construct viable arguments and critique the reasoning of others.**
4. **Model with mathematics.**
5. **Use appropriate tools strategically.**

##### Introduction

Before beginning the task, ask the class what they know about correlation. Remind them that the correlation coefficient, a measure of how closely two variables

are related, is a number between  $-1$  and  $1$ . If the values of both variables tend to increase (or if the values of both decrease), the two variables are positively correlated. If one variable tends to decrease as the other increases (or vice versa), the two variables are negatively correlated. If the values of the variables in both sets do not demonstrate a relationship, the variables are not correlated. Determining a relationship between two sets of data, especially from a scatter plot, may be subject to interpretation. The teacher will likely want to have students use a graphing calculator with statistical capabilities to do this task, determining ahead of time which features on the calculator are appropriate.

Lines of good fit may be found using paper-and-pencil techniques (such as writing the equation based on two points) or using a graphing calculator (either generating possible lines to use for guessing and checking or using the regression feature of the calculator to determine a particular function rule). Discuss correlation and causation with the group. Ask them at the end of the task to summarize television watching and test grades and if they believe there is a causal relationship. Have them defend their position based on statistical analysis.

**Materials**

- pencil
- graphing paper
- graphing calculator or statistical software package

**Prerequisites**

Students must have knowledge of writing linear equations based on two points and understand correlation.

**Time Required**

1 to 2 class periods.

- Students in Ms. Garth’s Algebra II class wanted to see if there are correlations between test scores and height and between test scores and time spent watching television. Before the students began collecting data, Ms. Garth asked them to predict what the data would reveal. Answer the following questions that Ms. Garth asked her class.

- Do you think students’ heights will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation?

***Answers may vary, but a possible answer could be: “I do not think there will be correlation between height and test grades, since it is not reasonable to think a person’s height affects their intelligence or effort level.”***

- Do you think the average number of hours students watch television per week will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation? Do watching TV and low test grades have a cause and effect relationship?

***Answers may vary, but a possible answer could be: “I think the average number of hours a student watches television will be negatively correlated with the student’s test grades. It is reasonable to think that the more TV you watch, the less time you spend studying, resulting in low test grades. However, it does not seem like these variables will be strongly correlated, since some people do not watch TV but do not spend time studying either. On the other hand, some students may watch a lot of TV and still study a lot.” Discuss correlation vs. causation with students. Give samples of variables that correlate and have them justify their argument.***

- The students then created a table in which they recorded each student’s height, average number of hours per week spent watching television (measured over a four-week period), and scores on two tests. Use the actual data collected by the students in Ms. Garth’s class, as shown in the table below, to answer the following questions.

<b>Student</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>
Height (inches)	60	65	51	76	66	72	59	58	70	67	65	71	58
TV Hrs/week	30	12	30	20	10	20	15	12	15	11	16	20	19
Test 1	60	80	65	85	100	78	75	95	75	90	90	80	75
Test 2	70	85	75	85	100	88	85	90	90	90	95	85	85

- a. Which pairs of variables seem to have a positive correlation? Explain.

*Test 1 scores and test 2 scores appear to be positively correlated. For the most part, student performance on both tests was fairly consistent, so students who did well on test 1 also did well on test 2, while those who did not do well on test 1 didn't do very well on test 2 either.*

- b. Which pairs of variables seem to have a negative correlation? Explain.

*Test 1 scores and hours per week watching television, and test 2 scores and hours per week watching television appear to be negatively correlated. In general, students who spent more time watching television had lower test scores than those who spent less time watching television.*

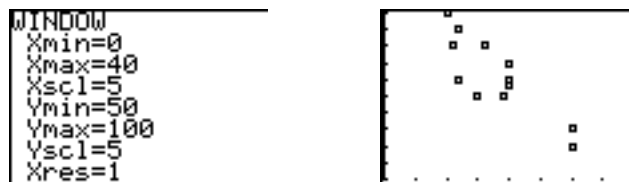
- c. Which pairs of variables seem to have no correlation? Explain.

*Height and hours per week watching television, test 1 scores and height, and test 2 scores and height seem to have no correlation. Height does not seem to be correlated with any of the other variables. That is, taller students do not seem to watch any more or less television or perform any better or worse on tests than shorter students.*

3. For each pair of variables listed below, create a scatter plot with the first variable shown on the y-axis and the second variable on the x-axis. Are the two variables correlated positively, correlated negatively, or not correlated? Determine whether each scatter plot suggests a linear trend.

- a. Score on test 1 versus hours watching television

*Scatter Plot:*



*Correlation? Negative. Linear Trend? Yes.*

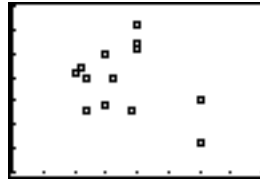
- b. Height versus hours watching television

**Scatter Plot:**

```

WINDOW
Xmin=0
Xmax=40
Xscl=5
Ymin=45
Ymax=80
Yscl=5
Xres=1

```



**Correlation? No correlation. Linear Trend? No.**

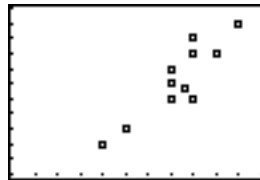
- c. Score on test 1 vs. score on test 2

**Scatter Plot:**

```

WINDOW
Xmin=50
Xmax=105
Xscl=5
Ymin=50
Ymax=105
Yscl=5
Xres=1

```



**Correlation? Positive. Linear Trend? Yes.**

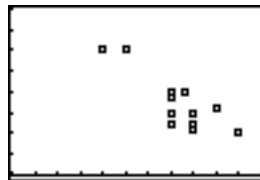
- d. Hours watching television versus score on test 2

**Scatter Plot:**

```

WINDOW
Xmin=50
Xmax=105
Xscl=5
Ymin=0
Ymax=40
Yscl=5
Xres=1

```



**Correlation? Negative. Linear Trend? Yes.**

4. Using the statistical functions of your graphing calculator, determine a line of good fit for each scatter plot that suggests a linear trend.

**Answers may vary slightly from the ones shown here.**

**Using linear regression and rounding to the hundredths place:**

- a. Score on Test 1 versus hours watching television:  $y = -1.43x + 105.98$
- b. Height versus hours watching television: *no linear trend*
- c. Score on test 1 versus score on test 2:  $y = 1.32x - 33.04$
- d. Hours watching television versus score on test 2:  $y = -0.72x + 79.64$

*Alternatively, using two points that appear to be close to a good representation of the trend in the data:*

*Data from score on test 1 versus hours spent watching television: (20, 78) and (11, 90)*

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{90 - 78}{11 - 20} = \frac{12}{-9} = \frac{-4}{3} = -1.33$$

$$y - y_1 = m(x - x_1)$$

$$y - 90 = \frac{-4}{3}(x - 11)$$

$$y - 90 = \frac{-4}{3}x + \frac{44}{3}$$

$$y - 90 = \frac{-4}{3}x + \frac{314}{3}$$

$$y = -1.33x + 104.67$$



## Guided Learning Task: TV/Test Grades

Name \_\_\_\_\_ Date \_\_\_\_\_

### **Common Core State Standards for Mathematical Practice**

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.

### **Common Core State Standards**

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6a** Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

**MCC9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**MCC9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient of a linear fit.

**MCC9-12.S.ID.9** Distinguish between correlation and causation.

1. Students in Ms. Garth's Algebra II class wanted to see if there are correlations between test scores and height and between test scores and time spent watching television. Before the students began collecting data, Ms. Garth asked them to predict what the data would reveal. Answer the following questions that Ms. Garth asked her class.
  - a. Do you think students' heights will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation?

b. Do you think the average number of hours students watch television per week will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation? Do watching TV and low test grades have a cause and effect relationship?

2. The students then created a table in which they recorded each student's height, average number of hours per week spent watching television (measured over a four-week period), and scores on two tests. Use the actual data collected by the students in Ms. Garth's class, as shown in the table below, to answer the following questions.

<b>Student</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>
Height (inches)	60	65	51	76	66	72	59	58	70	67	65	71	58
TV Hrs/week	30	12	30	20	10	20	15	12	15	11	16	20	19
Test 1	60	80	65	85	100	78	75	95	75	90	90	80	75
Test 2	70	85	75	85	100	88	85	90	90	90	95	85	85

- a. Which pairs of variables seem to have a positive correlation? Explain.
- b. Which pairs of variables seem to have a negative correlation? Explain.
- c. Which pairs of variables seem to have no correlation? Explain.
3. For each pair of variables listed below, create a scatter plot with the first variable shown on the  $y$ -axis and the second variable on the  $x$ -axis. Are the two variables correlated positively, correlated negatively, or not correlated? Determine whether each scatter plot suggests a linear trend.
- a. Score on test 1 versus hours watching television

- b. Height versus hours watching television
  - c. Score on test 1 versus score on test 2
  - d. Hours watching television versus score on test 2
4. Using the statistical functions of your graphing calculator, determine a line of good fit for each scatter plot that suggests a linear trend.

## APPENDIX C

### PERFORMANCE TASK: EQUAL SALARIES FOR EQUAL WORK?

#### Equal Salaries for Equal Work?

##### Mathematical Goals

- Represent data on a scatter plot
- Describe how two variables are related
- Informally assess the fit of a function by plotting and analyzing residuals
- Fit a linear function for a scatter plot that suggests a linear association

##### Common Core GPS

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6a** Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

**MCC9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**MCC9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient of a linear fit.

##### Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.

## **Introduction**

This task asks students to compare additive and multiplicative growth (represented by linear and exponential models) to make predictions and solve problems within the context of gender-based salary differences. In doing this task, students analyze data sets, create scatter plots, determine the most appropriate mathematical model, and justify their model selection.

This task provides a good example of how data points can appear to be linear over a relatively small domain, but how a different type of mathematical model might be more appropriate over a larger domain. This is an opportunity for students to discuss strengths and limitations of using mathematical functions to model real data. One discussion might arise as to whether other types of mathematical functions might sometimes be used for different types of data, perhaps leading students to look for patterns in data they might gather from sources like newspapers or books of world records.

Note that students will need to make a decision about the initial value representing the year. For example, it would be reasonable to assign the year 1984 (the first year in the table) as Year 0. The sample solutions below are based on this assumption.

## **Prerequisites**

Students must have knowledge of using the graphing calculator to create linear and exponential models and to analyze residuals. It is important that students understand how to assess the fit of a function to data and choose a function suggested by context.

## **Learning Targets**

When making statistical models, technology is valuable for varying assumptions, exploring consequences and comparing predictions with data. Students will interpret the correlation coefficient and show understanding of strengths and limitations of using mathematical functions to model real data.

## **Time Required**

1 class period

## **Materials**

Pencil and (graphing) paper; graphing calculator or statistical software package.

The data table shows the annual median earnings for female and male workers in the United States from 1984 to 2004. Use the data table to complete the task. Answer all questions in depth to show your understanding of the standards.

<b>Year</b>	<b>Women's median earnings (in dollars)</b>	<b>Men's median earnings (in dollars)</b>
1984	8675	17026
1985	9328	17779
1986	10016	18782
1987	10619	19818
1988	11096	20612
1989	11736	21376
1990	12250	21522
1991	12884	21857
1992	13527	21903
1993	13896	22443
1994	14323	23656
1995	15322	25018
1996	16028	25785
1997	16716	26843
1998	17716	28755
1999	18440	30079
2000	20267	30951
2001	20851	31364
2002	21429	31647
2003	22004	32048
2004	22256	32483

*Data provided by  
U.S. Census Bureau*

1. Create two scatter plots, one for women's median earnings over time and one for men's median earnings over time. Describe two things you notice about the scatter plots.

Each scatter plot below is graphed with the following window:

**Window:**

**Xmin=(-)2**

**X max=22**

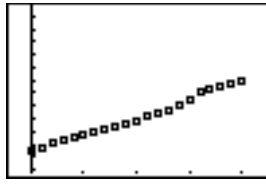
**Xscl=5**

**Ymin=4500**

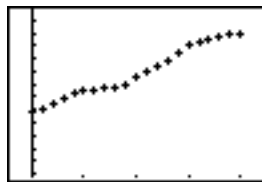
**Ymax=37000**

**Yscl=2500**

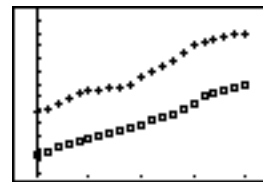
**Xres=1**



*Women's Data*



*Men's Data*



*Both Data Sets*

*Answers may vary.*

*Possible answers: From 1984 to 2004, median earnings for both men and women increased. In each of these years, men's median earnings were greater than women's median earnings.*

2. Terry and Tomas are trying to decide what type of model will most accurately represent the data. Terry thinks that a linear model might be most appropriate for each scatter plot. Help Terry find reasonable linear function rules for each scatter plot. Explain how you found these.

*Answers may vary*

*One solution*

*method:*

*To find a linear model of women's median earnings, use the starting earnings figure for women, \$8675, and the average rate of change of \$680 per year. (To find the average rate of change, find successive differences and then find the average of the successive differences.) The linear model is  $m(x) = 680x + 8675$ , where  $x$  represents years and  $m(x)$  gives the median earnings. To find a linear model of men's median earnings, use the starting earnings figure for men, \$17,026, and the average rate of change of \$773 per year. The linear model is  $m(x) = 773x + 17026$ , where  $x$  represents years and  $m(x)$  gives the median earnings.*

**Another solution method:**

*Using a graphing calculator to determine a regression line, women's median earnings could be represented by the function  $y = 703x + 8181$ .*

*Using a graphing calculator to determine a regression line, men's median earnings could be represented by the function  $y = 814x + 16709$ .*

3. Using the linear models, will women's annual median earnings ever equal those of men?  
Why or why not?

*Using the linear models created from the data provided, women's annual median earnings will never equal men's annual median earnings. The men's linear model has a larger y- intercept and a larger slope, meaning the men start out earning more money and also experience a faster rate of increase in earnings.*

4. Tomas thinks that an exponential model might be most appropriate for each scatter plot. Help Tomas find reasonable exponential function rules for each scatter plot. Explain how you found these.

*Answers may vary.*

**One solution method:**

*To find an exponential model of women's median earnings, use the starting income for women, \$8675, and the average quotient, 1.048. (To find the average quotient, find successive quotients then find the average of the successive quotients.) The exponential model is  $m(x) = 8675(1.048)^x$ , where  $x$  represents years and  $m(x)$  gives the median salary. To find an exponential model of men's median earnings, use the starting earnings figure for men, \$17,026, and the average quotient, 1.033. The exponential model is  $m(x) = 17026(1.033)^x$ , where  $x$  represents years and  $m(x)$  gives the median earnings.*

**Another solution method:**

*Calculating an exponential regression function on a graphing calculator, women's median earnings could be represented by the function  $y = 9087(1.049)^x$ .*

*Calculating an exponential regression function on a graphing calculator, men's median earnings could be represented by the function  $y = 17479(1.034)^x$ .*



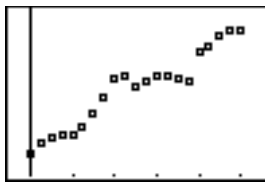
5. Using the exponential models, will women's annual median earnings ever equal those of men? Why or why not?

*Using the exponential models, women's annual median earnings will eventually equal those of men. The exponential model of men's earnings has a base of 1.034, and the exponential model of women's earnings has a base of 1.049. Since the women's model has a higher base, their earnings are increasing at a faster rate and will eventually surpass men's earnings. These functions can also be graphed to determine their intersection (45.7, 79533.88), demonstrating that at some point during the year 2029, women's annual median earnings will overtake men's annual median earnings.*

6. If you answered yes to either question 3 or question 5, use that model to determine the first year women will have higher median earnings than men. Explain how you found your answer.

*Using the exponential models, women's annual median earnings will eventually equal those of men. The exponential model of men's earnings has a base of 1.034, and the exponential model of women's earnings has a base of 1.049. Since the women's model has a higher base, their earnings are increasing at a faster rate and will eventually surpass men's earnings. These functions can also be graphed to determine their intersection (45.7, 79533.88), demonstrating that at some point during the year 2029, women's annual median earnings will overtake men's annual median earnings.*

7. For each year listed in the table, find the ratio of women's to men's annual median earnings expressed as a percentage. Use the data to create a scatter plot of percentage versus year. Based on this graph, do you think women's annual median earnings will ever equal those of men? Why or why not?



*The scatter plot has a positive correlation. This means that women's annual earnings are approaching those of men and (if the trend continues) will eventually catch up to men's annual median earnings.*

8. Considering the results of the scatter plot in question 7 above, do you think the linear model or exponential model makes more sense? Why?

*Answers will vary. Generally speaking, the exponential model makes more sense because the gap between men's earnings and women's earnings is decreasing, as shown in the percentage-versus-time scatter plot. This more closely represents the real situation. The linear model shows the gap widening — an inaccurate representation of what is actually happening.*

Data on earnings by gender provided by:

U.S. Census Bureau. "Table P-41. Work Experience—All Workers by Median Earnings and Sex: 1967 to 2005." *Historical Income Tables—People*.

[www.census.gov/hhes/www/income/histinc/p41ar.html](http://www.census.gov/hhes/www/income/histinc/p41ar.html). (Date retrieved: July 24, 2007.)

## **Performance Task: Equal Salaries for Equal Work!**

Name \_\_\_\_\_

Date \_\_\_\_\_

### **Common Core State Standards for Mathematical Practice**

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.

### **Common Core GPS**

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6a** Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

**MCC9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**MCC9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient of a linear fit.

**MCC9-12.S.ID.9** Distinguish between correlation and causation.

The data table shows the annual median earnings for female and male workers in the United States from 1984 to 2004. Use the data table to complete the task. Answer all questions in depth to show your understanding of the standards.

<b>Year</b>	<b>Women's median earnings (in dollars)</b>	<b>Men's median earnings (in dollars)</b>
1984	8675	17026
1985	9328	17779
1986	10016	18782
1987	10619	19818
1988	11096	20612
1989	11736	21376
1990	12250	21522
1991	12884	21857
1992	13527	21903
1993	13896	22443
1994	14323	23656
1995	15322	25018
1996	16028	25785
1997	16716	26843
1998	17716	28755
1999	18440	30079
2000	20267	30951
2001	20851	31364
2002	21429	31647
2003	22004	32048
2004	22256	32483

*Data provided by  
U.S. Census Bureau*

1. Create two scatter plots, one for women's median earnings over time and one for men's median earnings over time. Describe two things you notice about the scatter plots.

2. Terry and Tomas are trying to decide what type of model will most accurately represent the data. Terry thinks that a linear model might be most appropriate for each scatter plot. Help Terry find reasonable linear function rules for each scatter plot. Explain how you found these.
  
3. Using the linear models, will women's annual median earnings ever equal those of men?  
Why or why not?
  
  
  
  
  
  
  
  
  
  
4. Tomas thinks that an exponential model might be most appropriate for each scatter plot. Help Tomas find reasonable exponential function rules for each scatter plot. Explain how you found these.
  
  
  
  
  
  
  
  
  
  
5. Using the exponential models, will women's annual median earnings ever equal those of men? Why or why not?
  
  
  
  
  
  
  
  
  
  
6. If you answered yes to either question 3 or question 5, use that model to determine the first year women will have higher median earnings than men. Explain how you found your answer.
  
  
  
  
  
  
  
  
  
  
7. For each year listed in the table, find the ratio of women's to men's annual median earnings expressed as a percentage. Use the data to create a scatter plot of percentage versus year. Based on this graph, do you think women's annual median earnings will ever equal those of men? Why or why not?

8. Considering the results of the scatter plot in question 7 above, do you think the linear model or exponential model makes more sense? Why?

Data on earnings by gender provided by:

U.S. Census Bureau. "Table P-41. Work Experience—All Workers by Median Earnings and Sex: 1967 to 2005." *Historical Income Tables—People*.

[www.census.gov/hhes/www/income/histinc/p41ar.html](http://www.census.gov/hhes/www/income/histinc/p41ar.html). (Date retrieved: July 24, 2007.)

## APPENDIX D

### STUDENT WORKSHEET: LEARNING TASK 1

Name \_\_\_\_\_ Date \_\_\_\_\_

#### Learning Task: Simple Linear Regression

##### Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.

##### Common Core State Standards

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.

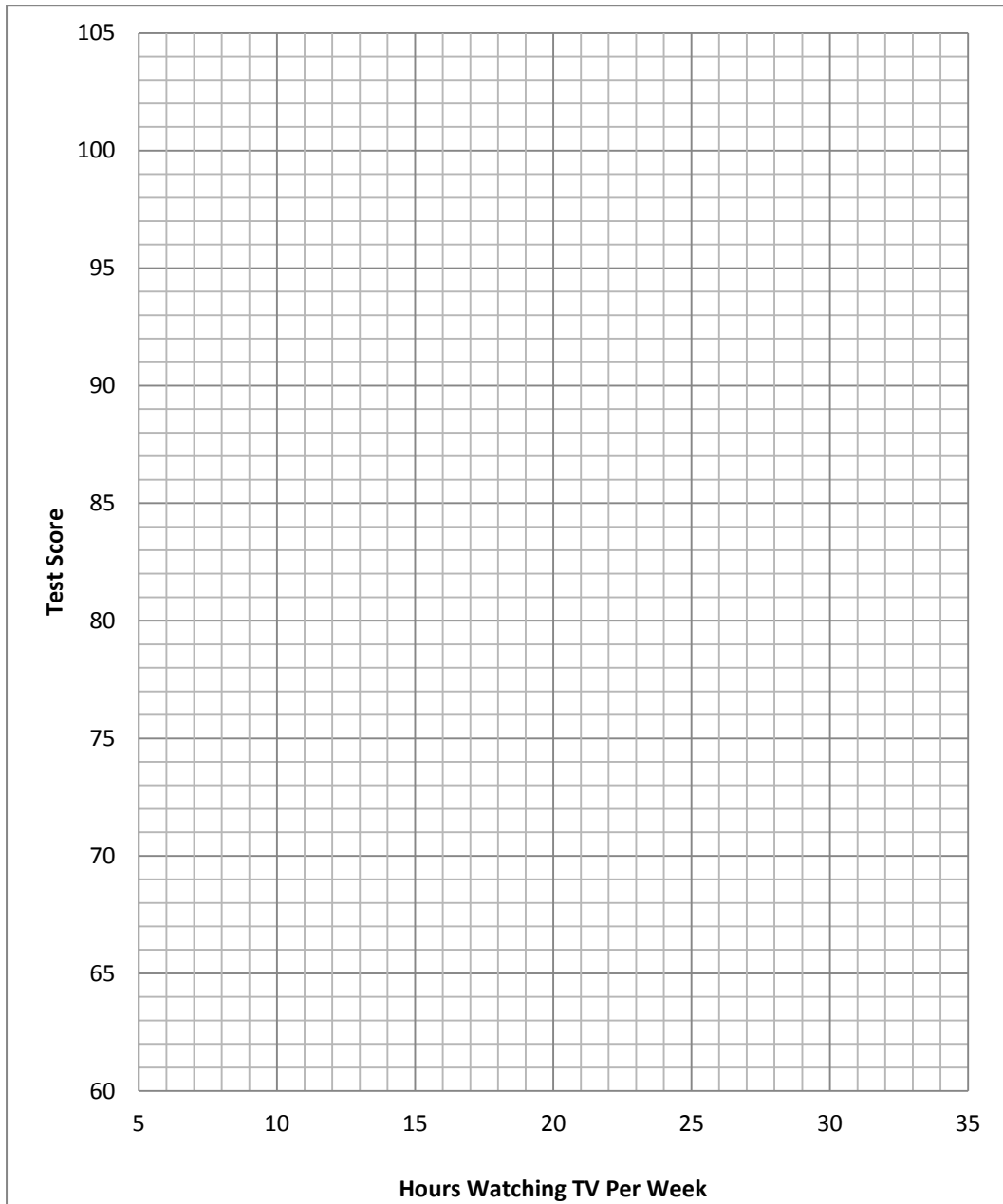
**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

#### Part I

3. Create a scatter plot of the following data set TV Time and Test Score. Plot the data points on the graph provided.

TV Time (Hours)	Test Score
30	70
12	85
30	75
20	85
10	100
20	88
15	85
12	90
15	90
11	90
16	95
20	85
19	85

4. Examine the scatter plot you created and visually determine a line of best-fit (or trend line). Draw your best-fit line on your scatter plot.





3. Now investigate the “goodness” of the fit. Measure the residual using the method from the Measuring Notes sheet. Repeat this for each point in the scatter plot.

4. Calculate the sum of your residuals.

Total error = \_\_\_\_\_

\* Class discussion before moving to Part II. \*

## **Part II**

5. Calculate the square of each residual. Find the sum of your squared residuals.

Total residual sum of squares = \_\_\_\_\_

6. As a class find the equation of the best-fit line.

7. Using the class best-fit line, what do you think your test score will be based on how many hours of TV you watch per week?

8. What can you conclude about the relationship between the number of hours of TV students watch per week and your test score?
10. **Extension:** Plot the residual plot of the data using your calculator. What can you conclude about the linear model for the data set? Is the linear model an appropriate representation of the data?

## APPENDIX E

### STUDENT WORKSHEET: LEARNING TASK 2

Name \_\_\_\_\_ Date \_\_\_\_\_

#### Learning Task: TV Watching Hours and Test Grades

##### **Common Core State Standards**

**MCC9-12.S.ID.6** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.

Interpret linear models

**MCC9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**MCC9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient of a linear fit.

**MCC9-12.S.ID.9** Distinguish between correlation and causation.

##### **Standards for Mathematical Practice**

- 1. Make sense of problems and persevere in solving them.**
- 2. Reason abstractly and quantitatively.**
- 3. Construct viable arguments and critique the reasoning of others.**
- 4. Model with mathematics.**
- 5. Use appropriate tools strategically.**

##### **Part I:**

1. Students in Ms. Garth's Algebra II class wanted to see if there are correlations between test scores and height and between test scores and time spent watching television. Before the students began collecting data, Ms. Garth asked them to predict what the data would reveal. Answer the following questions that Ms. Garth asked her class.
  - a. Do you think students' heights will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation?

- b. Do you think the average number of hours students watch television per week will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation? Do watching TV and low test grades have a cause and effect relationship?

2. The students then created a table in which they recorded each student's height, average number of hours per week spent watching television (measured over a four-week period), and scores on two tests. Use the actual data collected by the students in Ms. Garth's class, as shown in the table below, to answer the following questions.

<b>Student</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>
Height (inches)	60	65	51	76	66	72	59	58	70	67	65	71	58
TV hrs/week	30	12	30	20	10	20	15	12	15	11	16	20	19
Test 1	60	80	65	85	100	78	75	95	75	90	90	80	75
Test 2	70	85	75	85	100	88	85	90	90	90	95	85	85

- a. Which pairs of variables seem to have a positive correlation? Explain.

- b. Which pairs of variables seem to have a negative correlation? Explain.

- c. Which pairs of variables seem to have no correlation? Explain.

**Part II:**

3. For each pair of variables listed below, create a scatter plot with the first variable shown on the x-axis and the second variable on the y-axis. Are the two variables correlated positively, correlated negatively, or not correlated? What is their best-fit regression equation?

X	Y	Linear Correlation?	Linear Reg. Equation	Correlation coefficient
Height	TV Hrs			
TV Hrs	Test 1			
TV Hrs	Test 2			
Test 1	Test 2			

4. From Ms. Garth's class data, and the correlations calculated, which variables can you say affects test scores?
5. Maria has asked you to predict her Test 1 score and told you that she is 61 inches tall. What do you think her score will be and how much belief do you have in this?
6. On the other hand, Johnny says he does not know his height but he knows he watches 15 hours of TV per week. What do you think his score will be and how much belief do you have in this?
7. Lauren made the conclusion that watching TV causes lower test scores. Can she make this conclusion? If not, why not?

8. Jacob concluded that a student scoring well on test 1 result in a high score on test 2. Is this a valid conclusion?
  
9. What is your best-fit linear regression equation for TV Hours vs Test 2 scores? Interpret the slope of your equation in terms of the context.
  
10. **Extension Question:** Interpret the coefficient of determination ( $r^2$ ) for the variable pairs, TV Hours vs Test 1, TV Hours vs Test 2, and Test 1 vs Test 2.

## APPENDIX F

### STUDENT WORKSHEET: LEARNING TASK 3

Name \_\_\_\_\_ Date \_\_\_\_\_

#### Learning Task: U.S. Population Growth

The data table shows the population in the United States gathered by the U.S. Census Bureau from 1800 to 2010. Use the data table to complete the task. Answer all questions in depth to show your understanding of the standards.

Year	Population
1800	5,308,483
1810	7,239,881
1820	9,638,453
1830	12,866,020
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	39,818,449
1880	50,189,209
1890	62,947,714
1900	76,212,168
1910	92,228,496
1920	106,021,537
1930	122,775,046
1940	132,164,569
1950	150,697,361
1960	179,323,175
1970	203,302,031
1980	226,545,805
1990	248,709,873
2000	281,421,906
2010	308,745,538

1. Create a scatter plot of the given data set of population over time.

2. Fit a linear model to the data using your graphing calculator. Does this model seem to be the most appropriate model for this data set?
  
3. Stacy believes the data does not look linear, and thinks that an exponential model might be more appropriate for the data. Find the exponential model for the data. Do you think this is a more appropriate model?
  
4. Compare your models with the model in the example given in class. Do the models agree? Why do you think there is a difference?
  
5. Use both models to predict the population in year 2020. Which model makes sense in representing the growth of the US Population?
  
6. **Extension:** Fit two models to the population data. Changes in immigration laws occurred after World War 1 in 1918, so consider 1910 and 1920 as the splitting points of the data. Fit a linear and exponential model to both of the data subsets (1800-1910, and 1920-2010) and decide which model is more appropriate by graphing the residual plots.



## APPENDIX G

### STUDENT WORKSHEET: PERFORMANCE TASK

Name \_\_\_\_\_ Date \_\_\_\_\_

#### Performance Task: Teacher Salary

Use the data table to complete the task. Answer all questions in depth to show your understanding of the standards.

Year	John	Lucy	Mark
0	1,000	10,000	5,000
1	3,628	12,826	10,274
2	6,288	15,709	15,601
3	8,982	18,650	20,982
4	11,710	21,650	26,417
5	14,472	24,711	31,906
6	17,269	27,834	37,450
7	20,101	31,020	43,050
8	22,968	34,270	48,707
9	25,782	37,585	54,420
10	28,813	40,968	60,191
11	31,790	44,419	66,019
12	34,804	47,939	71,906
13	37,857	51,531	77,853
14	40,948	55,195	83,859
15	44,078	58,933	89,925
16	47,078	62,747	96,052
17	50,456	66,637	102,241
18	53,705	70,606	108,492
19	56,995	74,655	114,806
20	60,326	78,786	121,183
21	63,700	83,000	127,624
22	67,116	87,299	134,130
23	70,574	91,685	140,702
24	74,076	96,160	147,339
25	77,623	100,725	154,043

	John	Lucy	Mark
Initial Deposit (Principal)	1,000	10,000	5,000
Annual Interest Rate	1.25%	2%	1%
Contribution per pay period	100	100	200
Pay periods per year	26	26	26

1. Create three scatter plots of the data of savings amount throughout over age. Describe two things you notice about the scatter plots.
2. Terry and Tomas are trying to decide what type of model will most accurately represent the data. Terry thinks that a linear model might be most appropriate for each scatter plot. Help Terry find the linear regression equation for each of the data sets, and interpret the slopes of the equations with respect to the context.
3. Using the linear models, will John's savings ever equal Lucy's savings? If so, at what year in their savings will this occur?
4. Tomas thinks that an exponential model might be most appropriate for each scatter plot. Help Tomas find the exponential function for each scatter plot.
5. Using the exponential models, will John's savings ever equal Lucy's savings? If so, when will this happen? Explain how you found your answer.
6. Based on the scatter plot and models, which of the linear model or exponential model do you think best represents the data? Why do you think so?

7. Using the correlation coefficient, which of the savings accounts have the strongest correlation? Is this correlation positive, negative, or no correlation?
  
8. Using the data provided, consider the following scenario. Lucy decided to make an initial deposit of \$10,000 since she is starting at age 40, while John started his savings account at the age of 30. Also, Lucy decided to make the same contribution per pay period of \$100, but was able to get a higher annual interest rate on her savings account. She believes she would have more money than John when they retire at the age of 55. Is this belief true? Considering how long it took for John and Lucy to have equal savings, which savings plan do you think is better for saving more money?
  
9. **Extension:** Using residual plots check the appropriateness of the linear regression model for each of the three savings accounts.