Text Messaging is Time Consuming! What Gives?

Jeanie Gibson
Hutchison School
jgibson@hutchisonschool.org

Mary McNelis
St. Agnes Academy
mmcnelis@saa-sds.org

Anna Bargagliotti
Loyola Marymount University
abargagl@lmu.edu

Project-SET
www.project-set.com

Published: August 2013

Overview of Lesson
This activity allows students to perform a hands-on investigation in which they explore relationships between two variables. Students use real data that they download from the Census at School project website (http://www.amstat.org/censusatschool/). Each student works on a question of choice and uses their downloaded sample data accordingly. They then plot a scatterplot to view the relationship of interest and estimate the regression line. After each student has interpreted their regression line, they pool their findings as a class to explore the variability in the slopes, intercepts, and correlations that they found. As a class, they construct approximations to the sampling distributions for each of these statistics and use the sampling distributions to make assertions about the values of the population parameters.

GAISE Components
This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

Common Core State Standards for Mathematical Practice
1. Make sense of problems and persevere in solving them.
4. Model with mathematics
5. Use appropriate tools strategically.

Common Core State Standards Grade Level Content (High School)
S-IC. 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
S-ID. 6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
S-ID. 6a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data.

S-ID. 6c. Fit a linear function for a scatter plot that suggests a linear association.
S-ID. 7. Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.
S-ID. 8. Compute (using technology) and interpret the correlation coefficient of a linear fit.

NCTM Principles and Standards for School Mathematics
Data Analysis and Probability Standards for Grades 9-12

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:

- understand the meaning of measurement data and categorical data, of univariate and bivariate data, and of the term variable;
- understand histograms, parallel box plots, and scatterplots and use them to display data;
- compute basic statistics and understand the distinction between a statistic and a parameter.

Select and use appropriate statistical methods to analyze data:

- for bivariate measurement data, be able to display a scatterplot, describe its shape, and determine regression coefficients, regression equations, and correlation coefficients using technological tools;
- identify trends in bivariate data and find functions that model the data or transform the data so that they can be modeled.

Develop and evaluate inferences and predictions that are based on data:

- use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions;
- understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.

Prerequisites

Prior to completing this activity students should be able to identify the population and sample in any given situation involving random sampling. They should also understand that information from a sample is used to draw conclusions about the entire population. They should have a basic understanding of how to construct a dot plot and do an informal analysis of the shape, center and spread. They should be able, with the aid of software or a calculator, to find the least-squares regression line.

Learning Targets

After completing the activity, students will have an understanding of how to carry out a regression and create approximations to sampling distributions for the slope, intercept, and the correlation of a regression. They will be familiar with the process of taking repeated samples of the same size, and constructing dot plots.

Time Required

The individual activity is designed to take 30-45 minutes. After that is completed, the class activity is designed for another 30-45 minute time frame.

Materials Required
For this activity, in addition to the activity sheet (page 13) students will need a computer with Internet access and Excel. Students will also need some type of statistical software or graphing calculator capable of estimating regression equations.

Instructional Lesson Plan

The GAISE Statistical Problem-Solving Procedure

Many text messages are sent throughout each day. What factors could be related to the number of text messages one sends in a day? In this activity, we will explore the relationship between the number of text messages one sends in a day and several potential explanatory factors. The student will work individually and then share their results with the class. Once each student has obtained results, they will work together as a group to draw further conclusions.

I. Formulate Question(s)
Before beginning the activity, the teacher may wish to review the concepts of population, sample, population parameter and sample statistic, reinforcing student understanding of these foundational concepts for the lesson. For this lesson, the population of interest being discussed is the population of students that have filled out the Census at School survey. The teacher should define the dependent variable, y, in the context of this lesson and discuss what the parameters β and α represent. Also, teachers should ensure that students have had practice estimating the least-squares regression line using either statistical software or a graphing calculator. If students have not had adequate practice constructing and analyzing dot plots, this process should also be reviewed. The questions of interest for the activity are:

1. Does the number of hours you spend hanging out with friends in a day (x) increase or decrease with the number of text messages you send (y)?
2. Does the number of hours you spend doing homework in a day (x) increase or decrease with the number of text messages you send (y)?
3. Does the number of text messages you receive in a day (x) increase or decrease with the number of text messages you send (y)?

Students will be asked to choose the statistical question they would like to focus on.

***Note for Teachers*:** For small classes (less than 30 students) give students only two questions to choose from. Make sure that the class is evenly divided among the questions. Ideally, you would like to have at least 15/20 people or more working on the same question. If the class size is too small to achieve this number of people working on the same question, reduce the number of questions or have each student choose two out of the three questions to explore.

II. Design and Implement a Plan to Collect the Data
In order to collect the data to be used during this activity, students will be instructed to go to the Census at School website and download a random sample. Students will then have to briefly clean up the data in Excel in order to make it easily workable for the lesson.

Explain to students that to answer the question that they chose, they are going to download and work with a real data set. Instruct students to go to the following website in order to download the data:

http://www.amstat.org/censusatschool/

Have each student individually carry out the following steps:

1. Click on Random Sampler

2. Accept the Terms & Conditions

3. Select a sample size of 100 from All States and 9, 10, 11, and 12 grade levels. Include All Genders and All Years of data collection.

4. Download the data into Excel.

5. Open the data in Excel. Explain that they will see a large number of variables (labeled in each column).

6. Delete all the columns except for the following:
Gender, Text Messages Sent Yesterday, Text Messages Received Yesterday, Hanging out with Friends Hours, Doing Homework Hours

Here is a snapshot of a sample data set showing the first 50 subjects.

1	Gender	Text_Messages_Sent_Yesterday	Text_Messages_Received_Yesterday	Hanging_Out_With_Friends_Hours	Doing_Homework_Hours
2	Female	10	10	6	4
3	Male	80	80	45	0.005
4	Male	30	30	40	2
5	Male	1	1	1	
6	Female				
7	Female	5	5	20	5
8	Female	18	23	1	18
9	Male	80	84	53	7
10	Female	4	5	40	6
11	Male	30	28	8	8
12	Female	60	60	25	5
13	Male	100	100	15	2
14	Male	0	0	12	8
15	Male	6	6	10	14
16	Female	20	25	10	6
17	Male	0	0	54	0
18	Female				
19	Male	0	4	15	5
20	Female	500	400	7	1
21	Male				
22	Male	55	156	30	10
23	Male	50	50	35	2
24	Female	15	15	8	25
25	Female	10	10	10	11
26	Male	5	5	45	2
27	Male	3	3	14	1
28	Male	25	27	1	1

29	Male	43	44	2	13
30	Male	15	15	1	6
31	Male	378	380	42	7
32	Female	50	55	8	35
33	Female	0	0	49	2
34	Male	248	230	5	15
35	Male	5	5	35	4
36	Female	500	600	10	2
37	Female	67	34		
38	Male		20	12	16
39	Female			7	0
40	Male				
41	Female		0	20	1
42	Male		12	18	15
43	Female		1000	60	1
44	Female		200	30	15
45	Female	60	68	30	10
46	Male	150	200	10	2
47	Male	6	11	6	8
48	Male	150	250	60	9
49	Female	30	30	10	4
50	Male	150	152	4	6
51	Male	0	0	2	35

Is this a realistic value for the number of text messages received?

Figure 1. Sample data of 50 subjects sampled using the Census at School sampler.

***Note for Teachers*:** The class should discuss whether the data are realistic or whether a casewise deletion of certain observations might be necessary. In the case where data seem incorrect, the student or class may determine that deleting that observation from their data set is advised.

III. Analyze the Data

Using each of their individual data, students will then be asked to analyze it on their own. Each student will plot a scatterplot, examine whether there seems to be a linear relationship, point out any outliers, compute the regression line, and compute the correlation coefficient.

Here are snapshots of sample scatterplots showing sample data for each of the three questions.

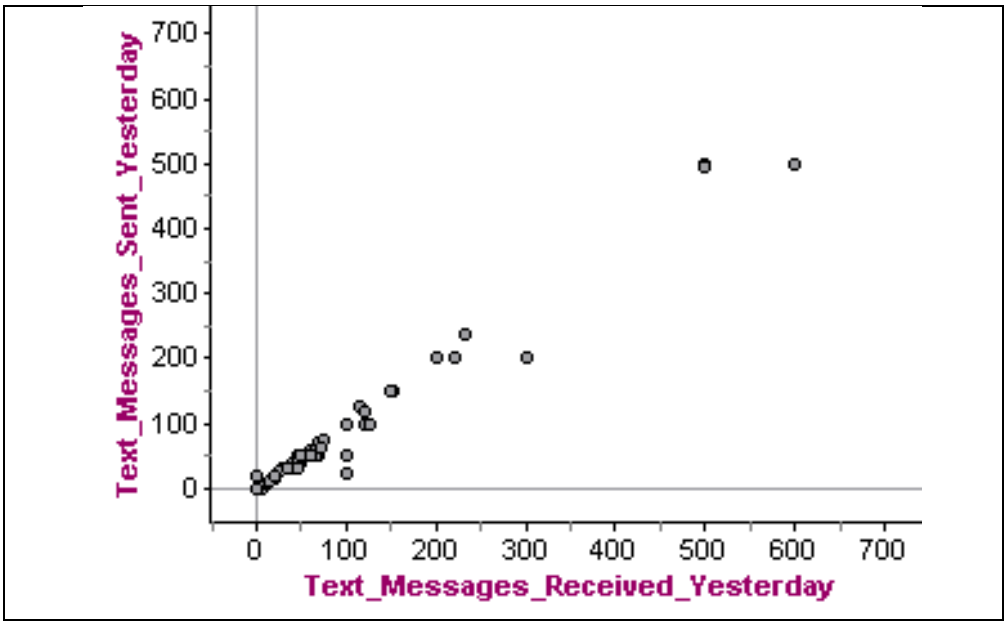


Figure 2. Scatterplot for sample of 50 subjects of texts received and texts sent.

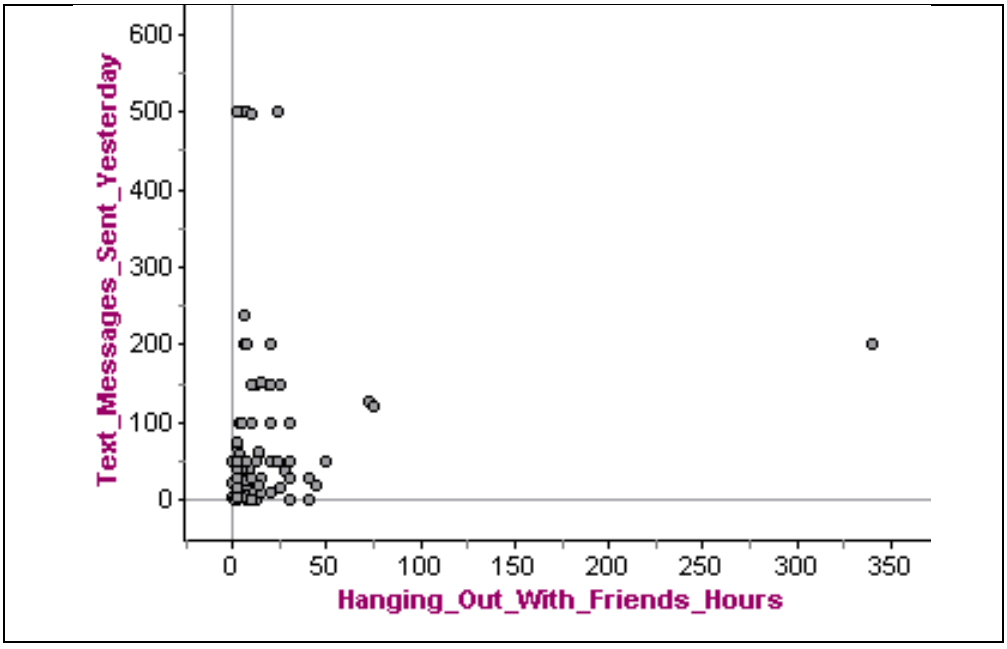


Figure 3. Scatterplot for sample of 50 subjects of hours per week spent with friends and texts sent.

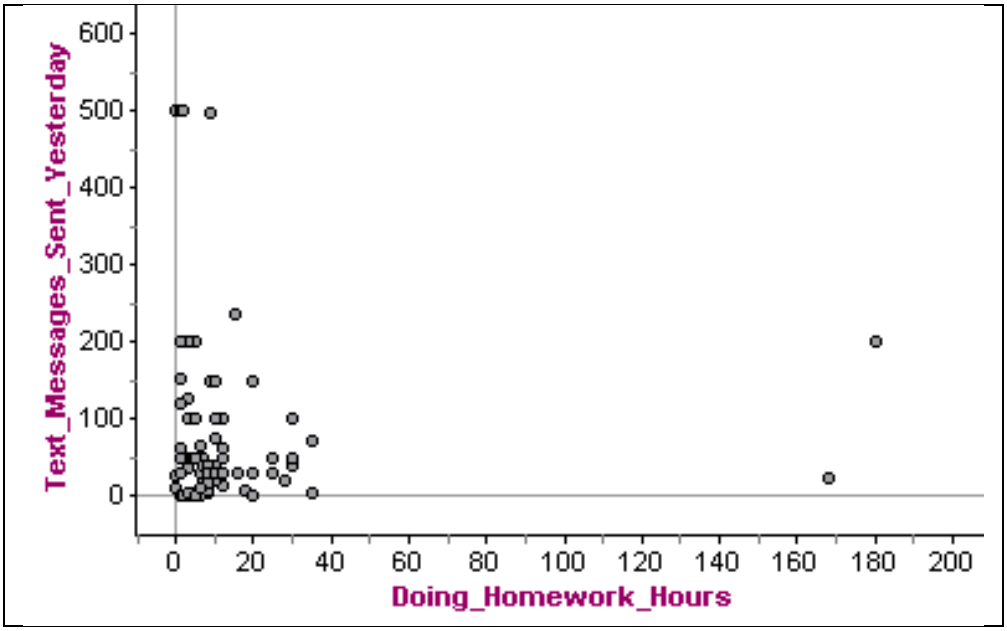


Figure 4. Scatterplot for sample of 50 subjects of hours spent on homework per week and texts sent.

Using technology students will estimate the regression lines:

If the student chose question (1) above, then they will estimate the following model:
Number of Text Messages Sent = $\alpha + \beta$(Hours Spent with Friends) + ε

If the student chose question (2) above, then they will estimate the following model:
Number of Text Messages Sent = $\alpha + \beta$(Hours Spent doing Homework) + ε

If the student chose question (3) above, then they will estimate the following model:
Number of Text Messages Sent = $\alpha + \beta$(Number of Text Messages Received) + ε

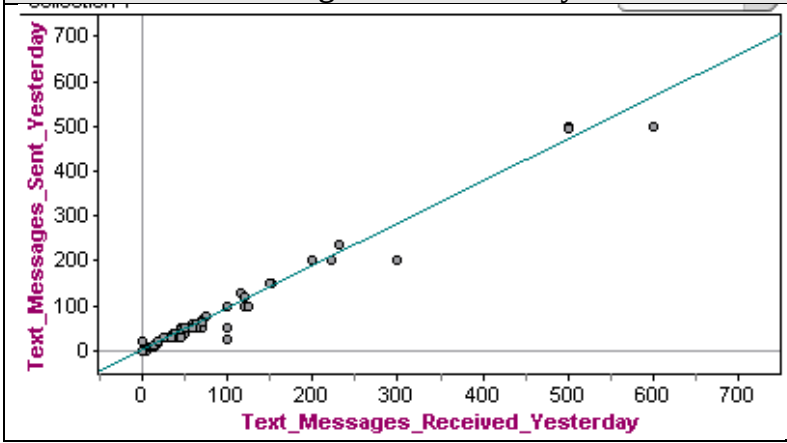
***Note for Teachers*:** The error term is included in this model to represent all other unknown factors that may contribute to the variation of Number of Text Messages Sent. When the model is actually estimated for a given sample data set, the error term no longer needs to be included.

This gives the least squares regression line as:

$$\hat{Y} = \_\_\_\_\_ X + \_\_\_\_\_$$

Here are three snapshots of sample scatterplots showing graphs and equations of the least squares regression lines for each of the three questions, along with the value of R^2 for each. The value of the correlation coefficient will be calculated using R^2.

Text Messages Sent Yesterday and Text Messages Received Yesterday

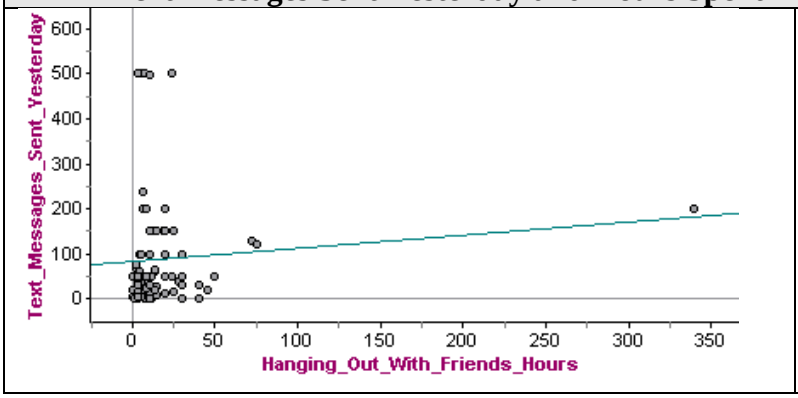


$\hat{Y} = .944X + .947$

$R^2 = .98$

Figure 5. Regression equation, correlation coefficient, and regression line drawn on the scatterplot of text messages received versus text messages sent for 50 subjects.

Text Messages Sent Yesterday and Hours Spent Hanging Out With Friends



$\hat{Y} = .293X + 80.02$

$R^2 = .01$

Figure 6. Regression equation, correlation coefficient, and regression line drawn on the scatterplot of hours hanging out with friends versus text messages sent for 50 subjects.

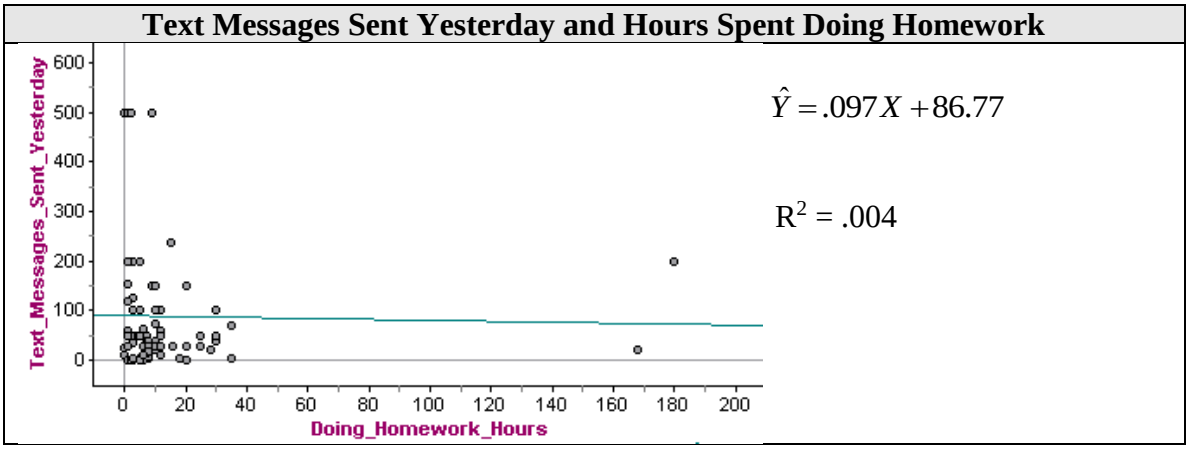


Figure 7. Regression equation, correlation coefficient, and regression line drawn on the scatterplot of hours doing homework versus text messages sent for 50 subjects

IV. Interpret the Results

Students will be asked to interpret the intercept and the slope in the context of their question. For example, an interpretation of the slope in Figure 5 would be that as the number of text messages received in a day increases by 1, the number of text messages sent in a day will on average increase by .944. The y-intercept may also be interpreted. In this example, the interpretation would be that when a person receives zero text messages in a day, a person on average still sends .947 texts.

Once their individual activities have been completed, they will work together as a class. Teachers will arrange the class into groups according to the questions they worked on. Together the groups will write in the equation of the line they found specifying the slope, intercept, and correlation coefficient they found. At this point, the class should have at minimum 15 to 20 slopes, intercepts, and correlations *for each* question. As a class, the students will plot these slopes, intercepts, and correlations in dot plots (a total of 6 dot plots will be created: three dot plots for each of the three questions). Students will then use these dot plots to make inferences about the population parameters for the slopes, intercepts, and correlations. Below are some example dot plots for the three different questions for the slope, intercept, and correlation coefficient:

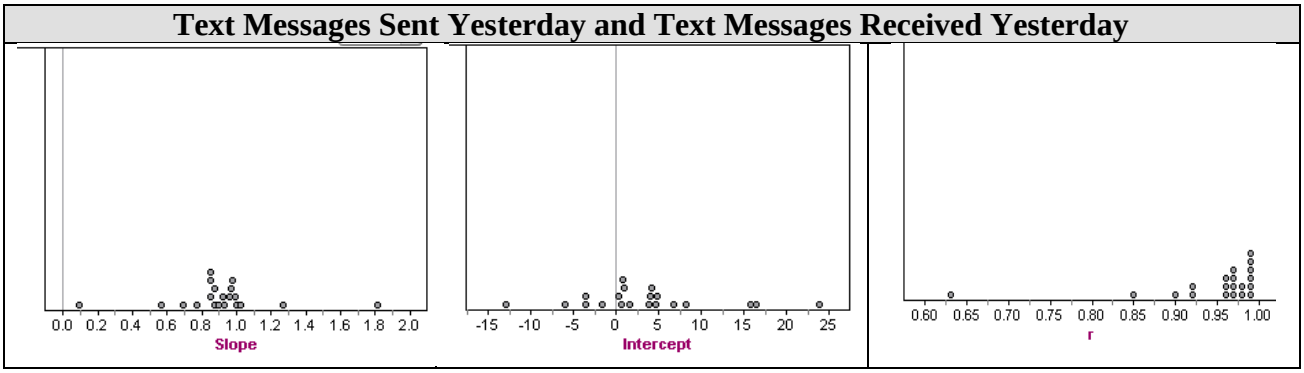


Figure 8. Dot plots for the slope, intercept, and correlation, *r*, for 20 samples of 50 subjects for text messages received versus text messages sent.

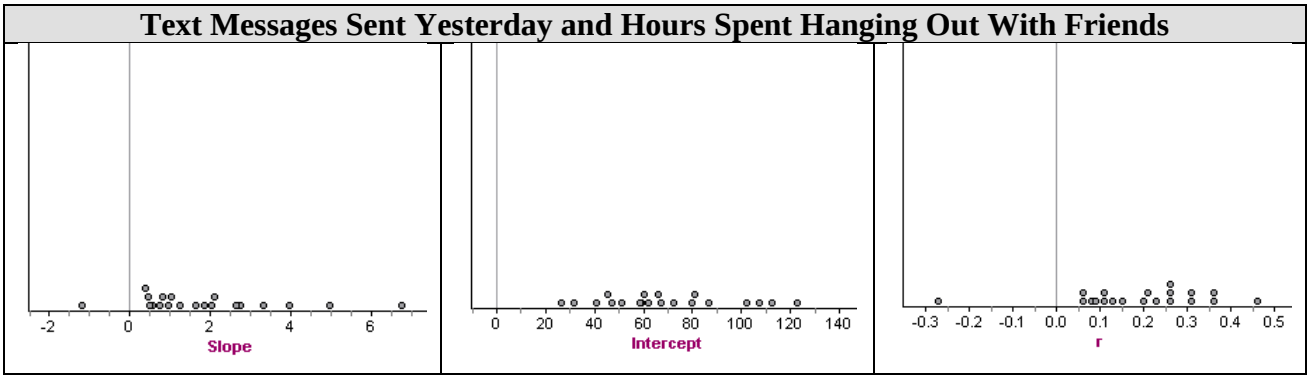


Figure 9. Dot plots for the slope, intercept, and correlation, *r*, for 20 samples of 50 subjects for hours spent hanging out with friends versus text messages sent.

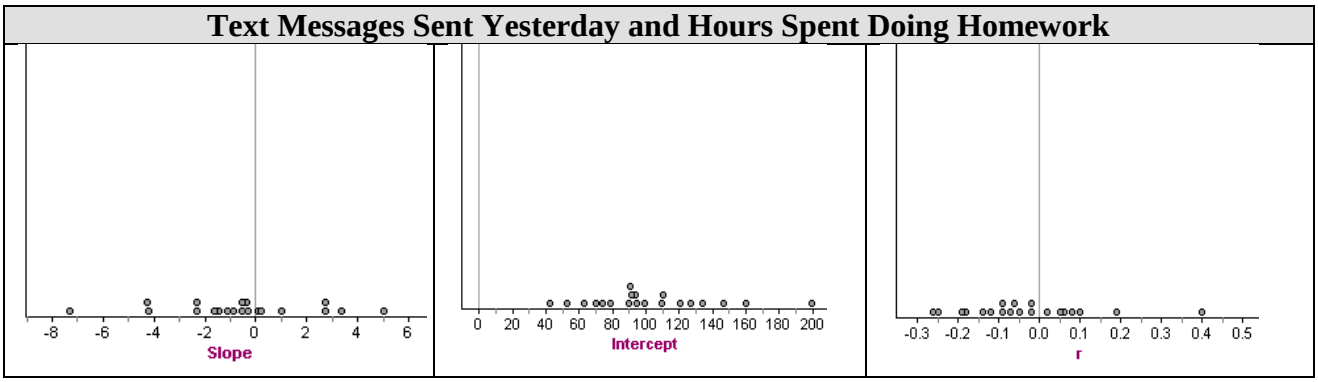


Figure 10. Dot plots for the slope, intercept, and correlation, *r*, for 20 samples of 50 subjects for hours spent doing homework versus text messages sent.

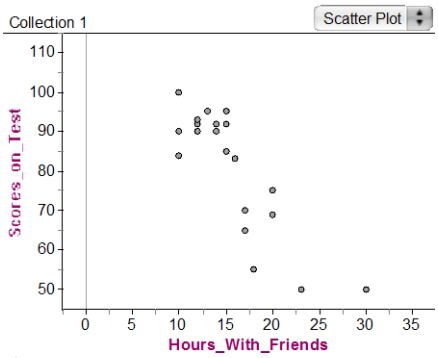
Here is a sample interpretation of the dot plots. Looking at Figure 8, one notices that most of the slope values on the dot plot are concentrated between .8 and 1 with only two values being far away from that interval (one value at 1.8 and one value at .1). The range of the values in the plot is 1.7. The most common slope appears to be .85. Given this approximation to the sampling distribution, one may guess that the slope of the regression line for the entire population be .9. The .9 appears to be near the center of the interval containing most of the values. In addition, .9 appears to be the mean of the values of the dot plot.

The dot plot of the intercept is less concentrated than the dot plot for the slope. Half the values sampled appear to be greater than about .5 while the other half are less than .5. The largest value is 24 and the smallest value is -14, thus giving a range of 38. The most common intercept is between .5 and 4. A best guess of the true population intercept would be .5. Because the values are split at about .5 and .5 appears to also be the mode of the dot plot.

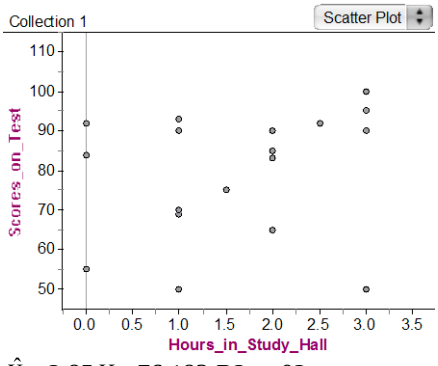
The dot plot for the correlation reveals that in most samples the correlation is very high. The distribution illustrates that except for one outlier value of .63, the correlation tends to be greater than .85. Moreover, the majority of the samples revealed a correlation greater than .95. Thus a good guess for the population correlation coefficient would be around .96. This would be a good guess because although the mode of the dot plot is .99, there are still some values that are a bit lower. A .96 correlation was found in three samples as well. The range of the dot plot is .36.

Assessment

At a large local high school, the principal wanted to ensure that her students would perform well on this year's standardized tests. As such, the principal came up with a list of factors that may negatively or positively impact test scores and aimed to prove it to the students while giving a practice test out of 100 points. A month before the practice test the principal asked students to fill out a survey asking them how many hours per week they hung out with their friends and how many hours per week they spent in study hall. Because the high school was very large, the principal only surveyed a sample of the students. The following two scatterplots provided show the results of the survey versus the student's scores on the practice exam.



$\hat{Y} = -2.69X + 122.87, R2 = .71$



$\hat{Y} = 2.85X + 76.183, R2 = .02$

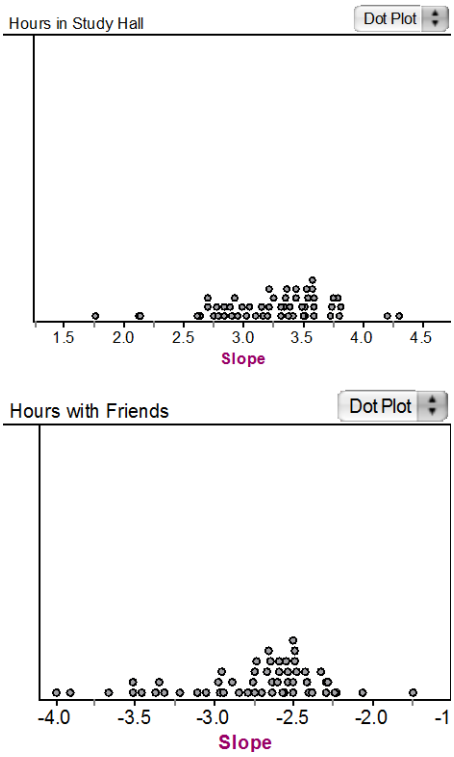
Based on these two scatterplots, answer the following questions.

1. Is there a positive or negative relationship between the hours a student spends with their friends and their test scores? Hours spent in study hall and their test scores?

2. On average, what would a student score if they spent zero hours per week hanging out with friends? In study hall?

3. On average, how many points on the test would a student increase/decrease if they spent 1 extra hour in study hall? Hanging out with friends?

When the students heard the results of the study, they asked the principal to look at different samples of students in the high school. To satisfy the students, the principal decided to randomly sample groups of 20 students at a time 15 more times. The following dot plots provide the summary of the results.



Based on the dot plots above, answer the following questions:

4. Should the students believe that the principal's decision to mandate an extra hour of study hall every week should increase their scores on the test? Explain.

5. Should the students try to decrease the number of hours they spent hanging out with friends before the test? Explain.

Contact Author for permission to use materials from this STEW lesson in a publication

Answers

1. There appears to be a negative linear relationship between the amount of time a student spends hanging out with their friends and their test scores. There does not seem to be a clear positive or negative relationship between the number of hours spent in study hall and the test scores.
2. On average, a student would score 122.87 on the test if they spent zero hours per week hanging out with friends. This y-intercept does not have a practical interpretation since there is no way to score more than 100 on the test. Also note that 0 is not within the range of the collected data values for hours spent with friends. On average, a student would score 76.183 on the test if they spent zero hours per week in study hall.
3. On average, a student's score will change by -2.69 points for every hour they spend hanging out with friends. On average, a student will increase 2.85 points on the test for every hour they spend in study hall.
4. The dot plot illustrates that all the sampled slopes are positive. This means that for every one of the 50 samples of 20 subjects sampled, the slope of the regression line was positive showing that as the number of hours of study hall increases, the scores on the test increase. In particular, the dot plot shows that the slopes tend to be for the most part between 2.6 and 3.6, meaning that on average scores would be raised between 2.6 and 3.6 for every hour extra spent in study hall.
5. The dot plot illustrates that all the sampled slopes are negative. This means that for every one of the 50 samples of 20 subjects sampled, the slope of the regression line was negative showing that as the number of hours of spent with friends increases, the scores on the test decrease. In particular, the dot plot shows that the slopes tend to be centered around -2.5, meaning that on average scores would change by about -2.5 for every hour extra spent in hanging out with friends.

Possible Extensions

1. Use the Census at School data to explore relationships between different pairs of variables.
2. Break up each sample by gender and repeat the activity for each gender separately. Compare whether there are differences between males' and females' text messaging experiences.

References

1. Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report, ASA, Franklin et al., ASA, 2007 http://www.amstat.org/education/gaise/

2. Census at School Project http://www.amstat.org/censusatschool/

3. Images purchased from: www.istockphoto.com

Text Messaging is Time Consuming! What Gives? Activity Sheet

Introduction

Many of us send lots of text messages throughout a day. What factors could be related to the number of text messages one sends in a day? In this activity, we will explore the relationship between the number of text messages one sends in a day and a few other potential explanatory factors. Each student will work individually and then share their results with the class. Once each student has obtained their results, we will work together as a group to draw further conclusions.

Choose one of the following questions to explore:

1. Does the number of hours you spend hanging out with friends in a day increase or decrease with the number of text messages you send?
2. Does the number of hours you spend doing homework in a day increase or decrease with the number of text messages you send?
3. Does the number of text messages you receive in a day increase or decrease with the number of text messages you send?

To answer the question you choose, you are going to download and work with a real data set. To download the data set, go to the following website: http://www.amstat.org/censusatschool/

Before we begin the individual activity, let's read about the Census at School project as a class on their home page. In this activity, we will be drawing random samples of students to analyze their results. In addition, at the end of the activity, you may also complete the survey.

You will each individually carry out the following steps:

1. Click on Random Sampler

2. Accept the Terms & Conditions

3. Select a sample size of 100 from All States and 9, 10, 11, and 12 grade levels. Include All Genders and All Years of data collection.

4. Download the data into Excel.

5. Open the data in Excel. You will see a large number of variables (labeled in each column).

6. Delete all the columns except for the following:

Gender, Text Messages Sent Yesterday, Text Messages Received Yesterday, Hanging out with Friends Hours, Doing Homework Hours

Contact Author for permission to use materials from this STEW lesson in a publication

7. Depending on which question you choose above, determine which is the dependent variable (y) and which is the independent variable (x)?

8. Depending on which question you choose above, plot a scatterplot to visually see the data. Does the relationship appear to be linear? Are there any outliers? What are some possible explanations for why there could be outliers? Should you eliminate the outliers in your data set? Why or why not?

9. Depending on which question you choose above, estimate the least squares regression line for your downloaded data. To carry out this step, you may use any technology available to you in the classroom. For example, you may use Excel, a graphing calculator, or Fathom. Use whatever technology you typically use in your classroom.

If you choose question (1) above, you will estimate the following model:
Number of Text Messages Sent = $\alpha + \beta$(Hours Spent with Friends) + ε

If you choose question (2) above, you will estimate the following model:
Number of Text Messages Sent = $\alpha + \beta$(Hours Spent doing Homework) + ε

If you choose question (3) above, you will estimate the following model:
Number of Text Messages Sent = $\alpha + \beta$(Number of Text Messages Received) + ε

This gives the least squares regression line as:

$$\hat{Y} = \_\_\_\_\_ X + \_\_\_\_\_$$

Interpret the slope (β) in the context of the question.

Interpret the y-intercept (α) in the context of the question.

Interpret the correlation coefficient in the context of the question.

STOP HERE AND WAIT FOR OTHERS TO FINISH 

Now, the class will be divided into three groups according to the question you chose to work on. Each group will complete the following table for their question.

10. Collect the regression line information found by each person in the class in a table:

Question Chosen: Number of Text Messages Sent vs._____

Student	Regression Line Equation	Slope	Intercept	Correlation
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

Contact Author for permission to use materials from this STEW lesson in a publication

11. Create a dot plot for the slopes found for each of the questions (i.e., three dot plots will be created – one for each question).

12. Look at the dot plot that pertains to your question. This dot plot represents an approximation to the **sampling distribution** of the slope of the regression line. What do you notice about the dot plot? What is the range of the slope? What seems to be the most common slope? If you had to guess what the slope of the regression line was for the entire population, what would you guess? Explain why.

Contact Author for permission to use materials from this STEW lesson in a publication

13. Repeat numbers 11 and 12 for the intercept.

Contact Author for permission to use materials from this STEW lesson in a publication

14. Repeat numbers 11 and 12 for the correlation.

Contact Author for permission to use materials from this STEW lesson in a publication