3.2 Correlation and Least Squares Regression AP Stat

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

How Strong is That Correlation??

**Description**

*Students explore the idea of correlation and least squares regression through baseball, chocolate, and human mortality.*

**Packet Contents**

* [Introduction](#intro)
* [Curriculum Alignment](#curriculum)
* [Objectives](#objectives)
* [Time and Location](#timeloc)
* [Teacher Materials](#teachermat)
* [Student Materials](#studentmat)
* [Safety](#safety)
* [Student Prior Knowledge](#studentpriorknowledge)
* [Teacher Preparation](#teacherprep)
* [Activities](#activities)
* [Assessment](#assessment)
* [Critical Vocabulary](#vocab)
* [Author Information](#authorinfo)

**Introduction**

During these lessons, students will use multiple data sets: candy, air pollution, baseball, trucks, and diamonds, to run least squares regression models, find residuals, and determine whether or not a regression model is appropriate for the relationship between the variables to be considered.

**Curriculum Alignment**

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

S-ID.9 Distinguish between correlation and causation.

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, quadratic, and exponential models.

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

N-Q.2 Define appropriate quantities for the purpose of descriptive modeling.

**Objectives**

The objectives list what students are expected to learn after completing the lesson plan.

* Students will be able to understand the basic properties of correlation, including how the correlation is influenced by outliers.
* Students will be able to distinguish correlation from causation.
* Students will be able to make predictions using regression lines, keeping in mind the dangers of extrapolation.
* Students will be able to calculate and interpret a residual.
* Students will be able to interpret the slope and y-intercept of a regression line.
* Students will be able to interpret the standard deviation of the residuals and r square and use these values to assess how well a least squares regression line models the relationship between two variables.
* Students will be able to find the slope and y intercept of the least squares regression line from the means and standard deviations of x and y and their correlation.
* Students will be able to describe how the least-squares regression line, standard deviation of the residuals, and r-squared are influenced by outliers.

**Time & Location**

2-3 90-minute class periods

Classroom

**Teacher Materials**

**The Practice of Statistics Sixth Edition -Starnes and Tabor**

**Bowl of wrapped candies**

**Calculators**

**Projector**

**Laptop**

**Doc Cam**

**Computers, if possible**

**Whiteboards**

**Whiteboard markers**

**Student Materials**

Computers

Calculators

Notebooks

Pencils

Whiteboard markers

**Student Prior Knowledge**

**Students should be able to create scatterplots, interpret basic correlations, recognize linear equations, and describe what a y-intercept and slope are in terms of a linear equation.**

**Teacher Preparations**

Students should be arranged in pairs, but be near enough to other pairs that they can discuss their thinking with each other when instructed. Computers should either be out on the desk warming up, or easy for students to get when they walk in and not take very long to warm-up, or else the first part of the day might take too much time. If your computers are not warmed up or quick to warm up, consider doing the journal activity together as a class. Also, double check that the applet doesn’t work well on mobile devices. The technology may have caught up with the times by the time you are implementing this lesson.

**Activities**

**Day 1**

10 min: Journal: Have students go to <http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html> when they enter class (might only work on computers). Have the instructions from the Correlation and Regression Applet Activity on p.163 either projected on the board, handed to the students, or instruct the students to turn to p.163 (section 3.1: Correlation and Regression applet activity) and follow the instructions.

10 min: Notes on “Cautions about Correlation” (p.163-166). Be sure to discuss the importance of the “correlation is not causation” piece!!

10 min: Walk students through how to calculate the correlation coefficient by hand (p.166). (Tell them, “never fear, there is a way to do this on the calculator, but it is important to know what is happening behind all of the calculator’s genius” or something like that).

10 min: Have them calculate the correlation coefficient for one of their scatterplots/data sets from the day before (this can be done individually, in pairs, or in a group…I tend to prefer pairs).

5 min: Then ask them to interpret the relationship based on that correlation coefficient (they learned this yesterday). Have them share with the other pair at their group what their interpretation is and how they can justify their answer.

10 min: Use p.176-178 to introduce and take notes on a regression lines, prediction, and extrapolation…make sure to use the first paragraph of Section 3.2 to help students compare and contrast regressions and correlations. Namely, that regression depends on having an explanatory variable and a response variable, but for correlation, it doesn’t matter which one is which.

5 min: Provide a regression line of the data sets that they chose (included in the drive folder) and have them make predictions. Make sure to make the connection to how this is actual data from a research lab in RTP that people are paid to analyze.

10 min: Then have each pair/group bring their regression line to the doc cam and have them discuss it with the class, explaining the form, direction, strength, outliers, how the correlation coefficient supports the direction and strength, and how valid their predictions are.

5 min: Individual practice with prediction using candy grab example (p.178-179). Tip: provide them the instructions separate from the book, because the book gives them the solution and you’ll want to separate that from their thinking time.

10 min: Residuals notes (p.179-180), introduction to calculating and interpreting residuals

5 min: Exit Ticket is the “Check Your Understanding” on p.181

**Day 2**

10 min: Journal: Project one of the screenshots from the air pollution data sheet in the Google Drive folder on the board and ask students to make a prediction, calculate a certain residual, and interpret the strength of the regression based on the correlation coefficient (which are basically the questions in the exit ticket from the day before)

10 min: Notes on Interpreting a regression line (p.181-182). MAKE SURE TO HIT THAT AP EXAM TIP! Perhaps encourage students to keep a running page in their notebook for AP Exam Tips.

5 min: Interpret their regression lines from the calculations they did the day before.

10 min: Discuss their interpretations at the doc cam, make sure that students are really grasping the slope interpretation (change in y over change in x) because that one is commonly mixed up.

5 min: Individual practice with Candy Grab example (p.182). Again, provide the instructions to students without allowing them to access the book, since they provide the solutions in the book. Allow students time to think before revealing the solutions and discussing any confusions.

5 min: Quick assessment: Check your understanding on p.182. Perhaps create a Socrative/Google forms question with it, or some other quick online assessment tool so they can see where they stand.

10 min: Discuss the least squares regression line and spend time on the technology corner (p.184), teaching students how to make the calculator do all of the heavy lifting for them. Fun fact: statisticians chose to do least squares regression instead of least absolute value regression or some other positive distance regression because the derivations of the other formulas in statistics were easier to calculate with squares. The math was easier! All mathematicians are lazy.

5 min: Take notes on residual plots and how to use them to determine how well a line fits the data.

10 min: Create residual plots for the data that they have already manipulated (continual application of the learning).

10 min: Take notes on r^2 and discuss. Be sure to harp on the language: “This tells us that \_\_\_% of the variance is accounted for by the least squares regression line.”

10 min: Exit Ticket: “Grabbing candy, again” example (p.191).

**Day 3**

10 min: Have students read to themselves about computer regression outputs (p.192) and then create one on their own of the regression information they have already found for all of their data so far.

10 min: Show students the different computer outputs from the air pollution data, as well as the book’s examples (p.193, 194) and have them write and interpret regression lines from the outputs.

20 min: Dig into where all of these numbers that we have already calculated come from by calculating them by hand using the five number summaries (p.194-196). Make sure to talk about how the regression is related to the means.

15 min: Read and discuss the cautions about correlation (p.197-201).

10 min: This is the last thing that we will do with these data sets. Go back to our work and make interpretations based on the three (some books have four) limitations of correlation and regression. In this exercise, students might say, “the points on the scatterplot are too random to really be considered a linear relationship, so the regression and correlation don’t do a good job modeling this data.” Or “since correlation and regression lines are not resistant, removing this one point would make the model a much better fit for this data.” Or “since association does not imply causation, even if the correlation coefficient was really close to one, we still could not say that pollution causes death.” **Push your students to write an interpretation for each context they used and for each of the limitations.**

10 min: Dig into a free response question on linear regression from the released AP Questions (2011B #6, 2011 #5, 2010B #6, 2007B #6, 2006 #2, 2005 #3, 2002 #4, 1999 #1, 1998 #4, 1997 #6)

15 min: End on a multiple choice quick quiz (attached in the Assessment section)

**Assessment**

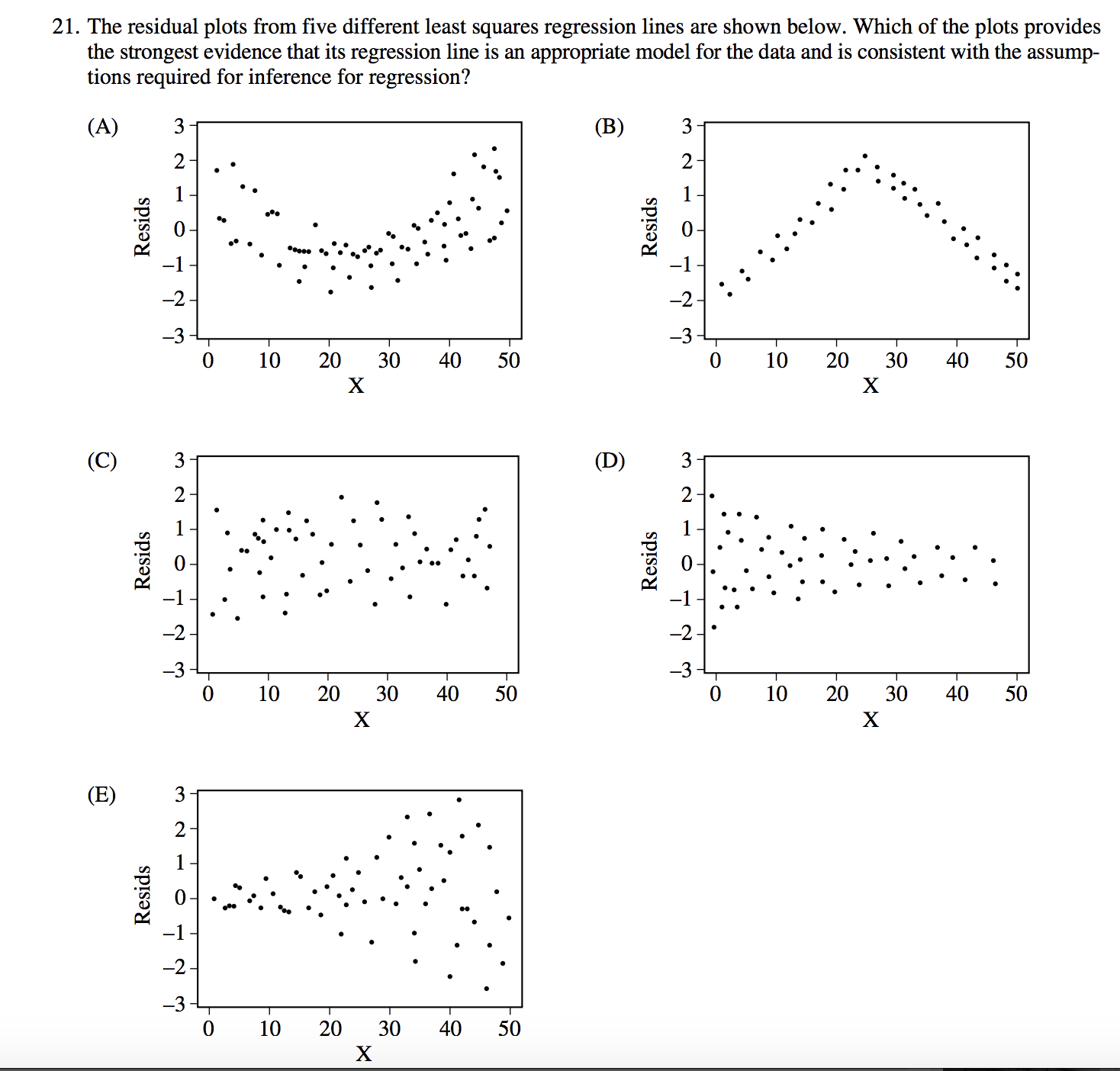
**For book activities, all answer keys can be found in the teacher edition of the book and many are found right there in the book for the students to use.**

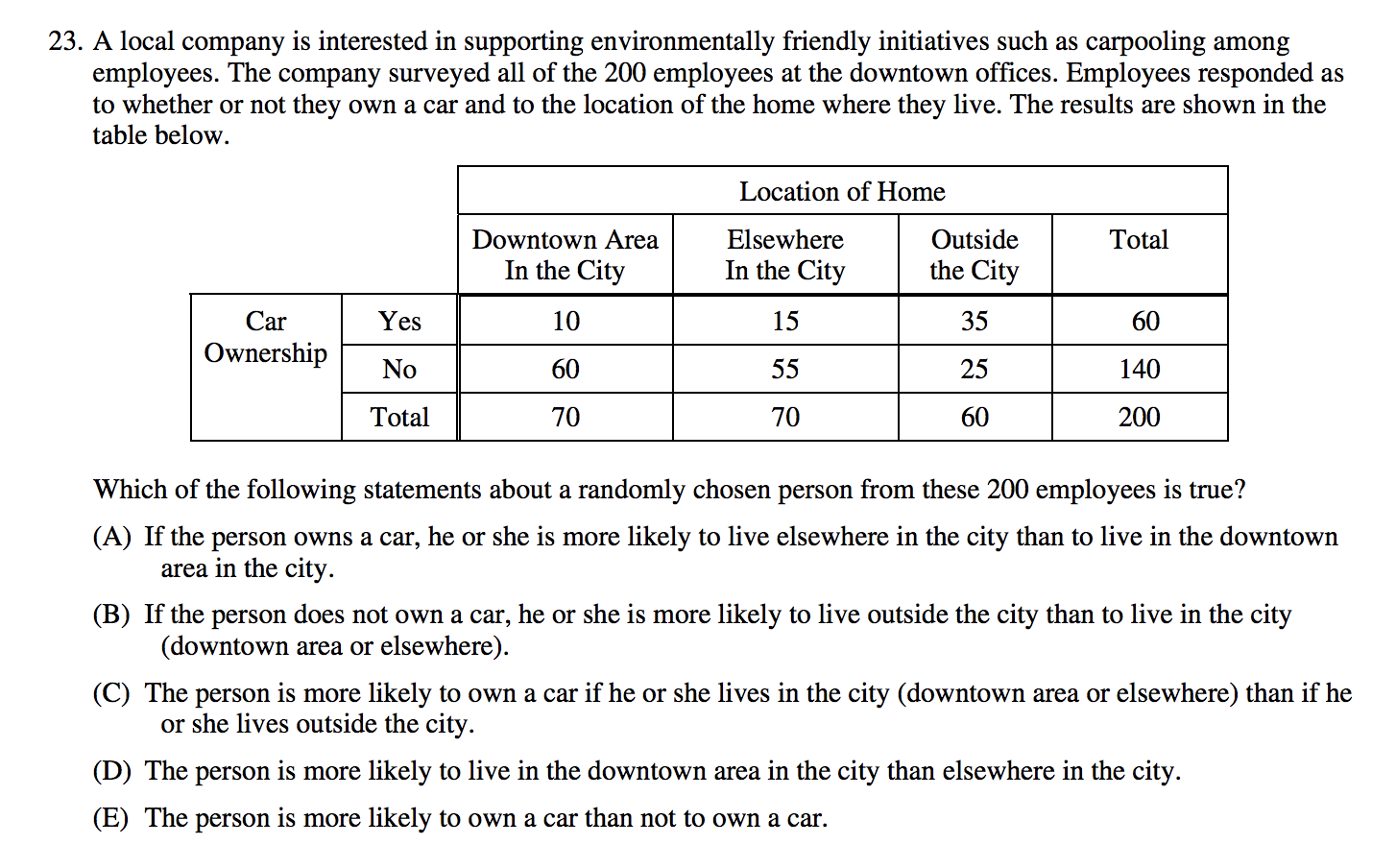
**All free response keys and rubrics can be found on the AP Central website.**

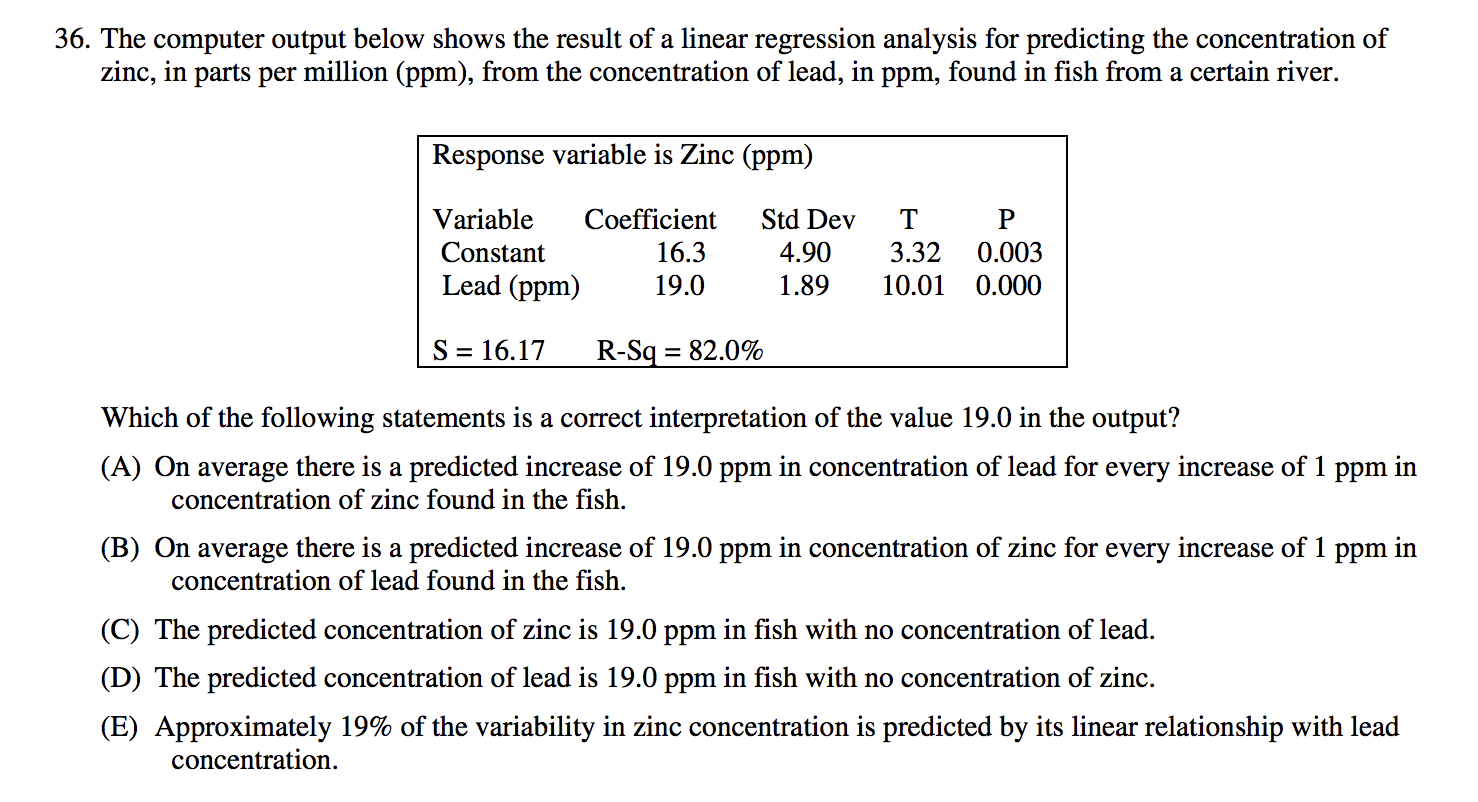
**My keys can be found both at the end of this document and at the end of the document with the data.**

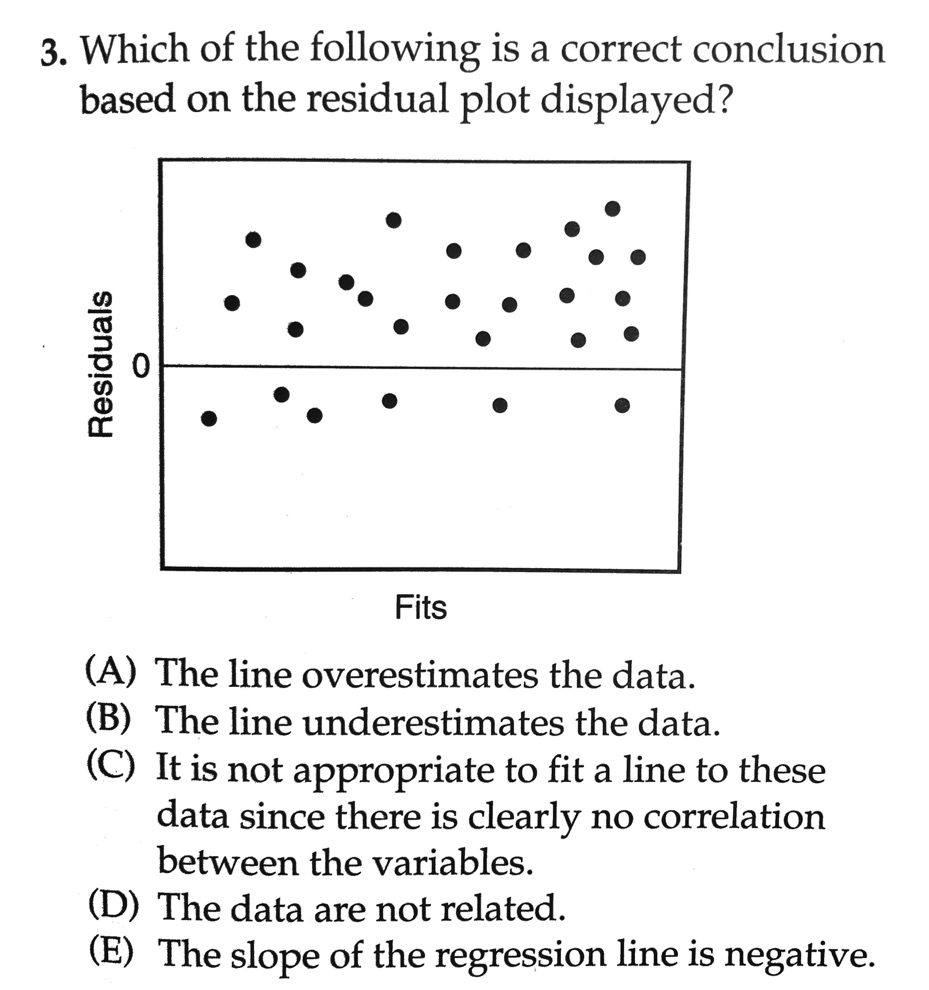
**Multiple Choice Quick Quiz: This should take students 10 minutes, if following AP Stats timing, and I often have them grade each other’s quizzes for immediate feedback to them and ability to discuss with each other.**

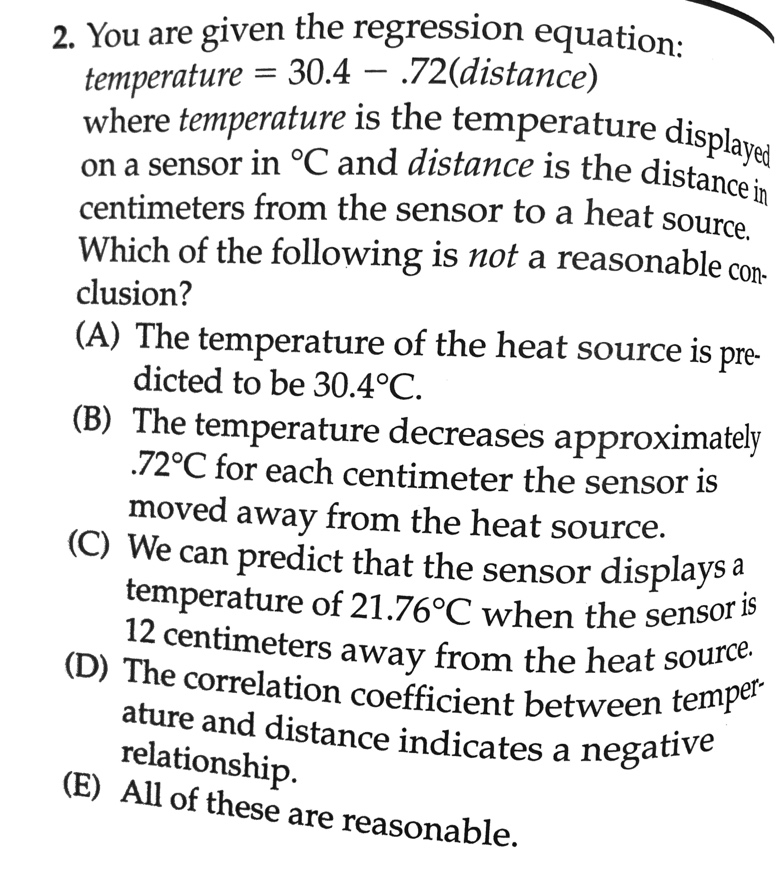
Stats Linear Regression and Residuals Quiz











Stats Linear Regression and Residuals Answer Key

21. C

23. A

36. B

3. B

2. E

**Critical Vocabulary**

Regression line: a line that describes how a response variable y changes as an explanatory variable x changes (expressed )

Extrapolation: use of a regression line for prediction far outside the interval of x values used to obtain the line

Residual: the difference between the actual value of y and the value of y predicted by the regression line (residual=actual y – predicted y = y - )

Y-intercept: the predicted value of y when x=0 ()

Slope: the amount by which the predicted value of y changes when x increases by 1 unit ()

Least squares regression line: line that makes the sum of the squared residuals as small as possible

Residual plot: scatterplot that displays the residuals on the vertical axis and the explanatory variable on the horizontal axis

Standard deviation of the residuals (s): measures the size of a typical residual, aka the typical distance between the actual y values and the predicted y values

**Author Information**

**Margaret Borden:**

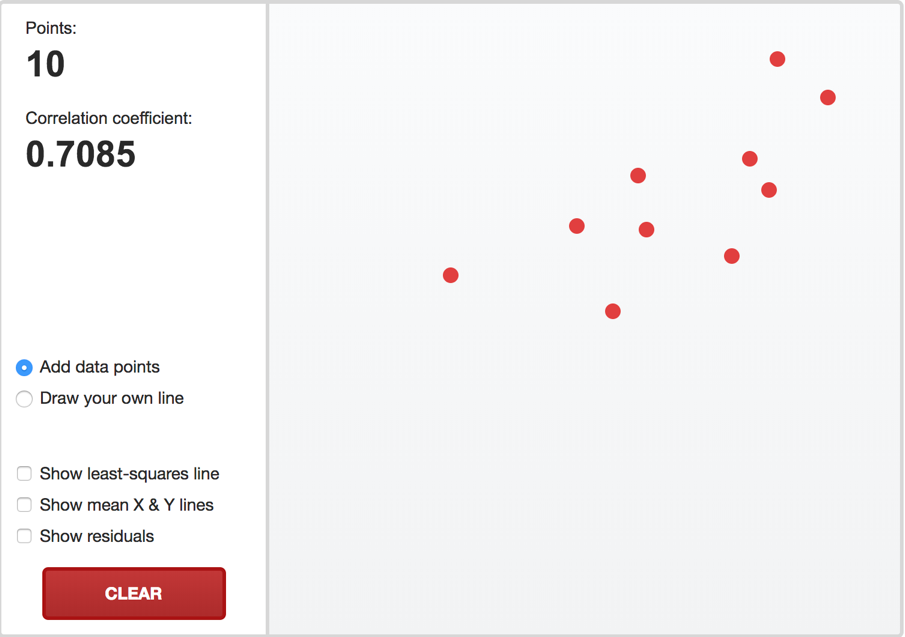
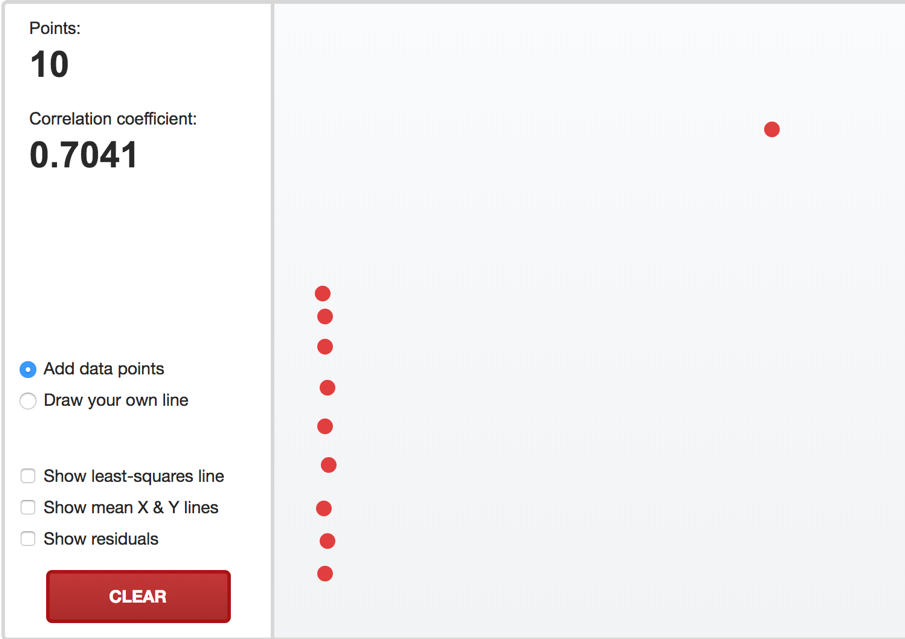
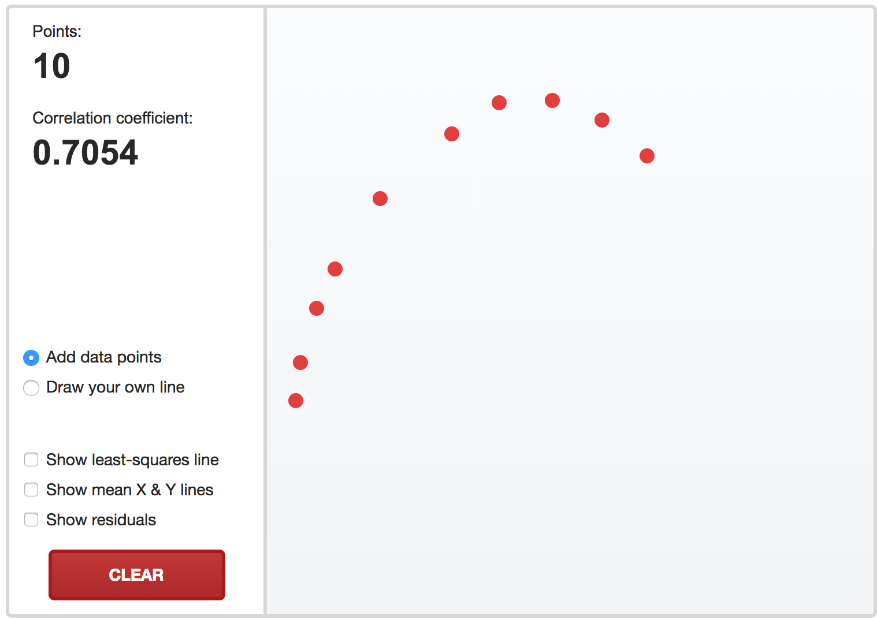
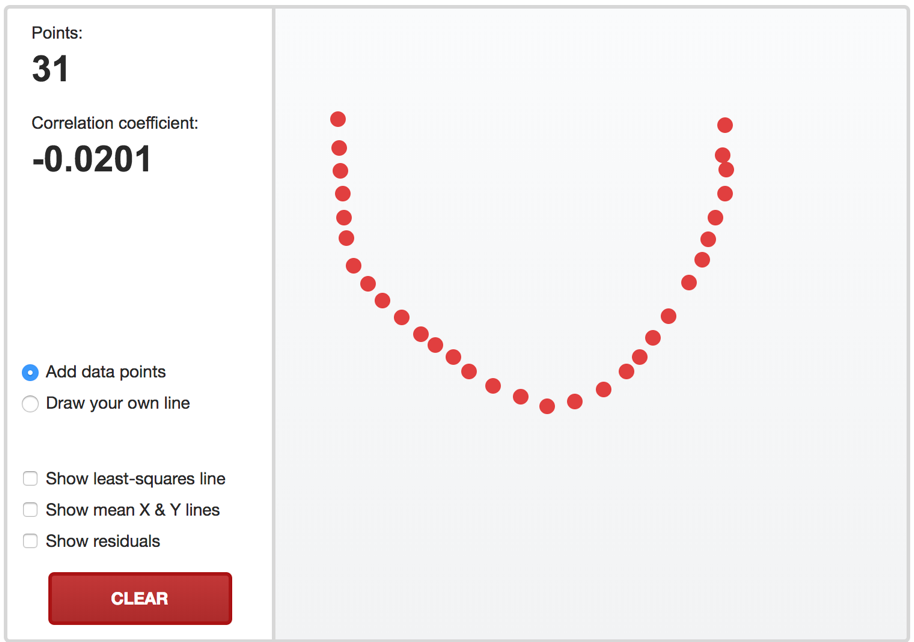
* **Knightdale High School of Collaborative Design, Wake County Public School System, Raleigh**
* **9-12 Math**
* **Second year**
* [mcleak@ncsu.edu](mailto:mcleak@ncsu.edu)**;** [mborden@wcpss.net](mailto:mborden@wcpss.net)

**Richard Smith:**

* **SAMSI and UNC Chapel Hill Department of Statistics and Operations Research**
* **Statistical Research**
* **Mark L. Reed III Distinguished Professor, Director of SAMSI, http://www.unc.edu/~rls/**
* [rls@samsi.info](mailto:rls@samsi.info)

Key

Correlation and Regression applet

1. 1.000 because there are only two points, so they can make a straight line
2. 
3. 
4. 
5. Absolutely nothing
6. 

All you know is that it doesn’t have a strong linear association