Visualizing and Understanding Regression and Correlation Using Dynamic Software

By Sheldon P. Gordon[1] and Florence S. Gordon[2]

[1] Farmingdale State College of SUNY, Farmingdale, NY 11731, USA
gordonsp@farmingdale.edu
[2] New York Institute of Technology, Old Westbury, NY, 11568, USA
fgordon@nyit.edu

This article illustrates ways that dynamic software using some sophisticated techniques in Excel can be used to demonstrate fundamental ideas related to regression and correlation analysis to increase student understanding of the concepts and methods in elementary statistics courses and in courses at the college algebra/precalculus level that stress ideas on curve fitting techniques.

1 INTRODUCTION

One of the standard topics in introductory statistics is linear regression and correlation analysis. Regression is the process of finding the best line, in the least squares sense, to fit a set of data that falls into a roughly linear pattern; correlation is a way to measure how well the regression line actually fits the data.

Unfortunately, for many students in introductory statistics, the ideas on regression lines seem to have nothing to do with anything they have seen before in mathematics. Their vague recollection from algebra classes is that lines have something to do with finding an equation based on two points using a quantity called the slope that they often never really understood. The fact that regression involves dozens, if not hundreds, of points and that the line they are expected to produce does not necessarily pass through any of them seems like a totally different concept.

Moreover, the calculations involved in finding the equation of the regression line and the value of the correlation coefficient are so involved that it is almost essential to utilize technology. All graphing calculators contain routines for constructing the regression line for any set of *x-y* data and the value of the associated correlation coefficient, as do statistical software packages and spreadsheets such as Excel. It is this capability that makes it reasonable to bring these topics to the forefront of introductory statistics courses, as well as courses in college algebra and precalculus that focus on curve-fitting as a way to illustrate the use of various families of functions. At the same time, the techniques all too often become reduced to button-pushing routines and far too many students come away with very little in the way of basic understanding of the underlying fundamental concepts. In particular, they don't understand the significance of:
- the variation in the results due to the variations between different samples;
- the effects of sample size on the results;
- the effects of changes in the sample data or the sample statistics on the results.

Yet, these are the fundamental issues that lie at the heart of every statistical technique, not just regression and correlation. Gaining a solid understanding of these ideas requires the use of dynamic software to bring the ideas to life and so make a far stronger impact on the students, as we demonstrate in this article.

2 BACKGROUND

Although in most mathematicians' minds, the algebra-precalculus-calculus and beyond track is the dominant focus for the mathematics curriculum, at least in the United States, introductory statistics courses have eclipsed calculus in terms of the total number of students taking courses. At the college and university level, in 2010, some 449,000 students took introductory statistics courses offered by mathematics departments and statistics departments (Blair et al., 2012). This total was growing at an annual rate of slightly over 7%. In addition, many times more students take introductory statistics courses offered by other departments that prefer to give their own versions of the course under names such as biostatistics, econometrics (for economics), psychometrics (for psychology), social statistics (for sociology), and so on. Admittedly, probably a good 90% of the content of these courses overlap.

In addition, statistics is rapidly growing in popularity as an elective course in American high schools. In 2016, over 206,000 students took the national final exam in the Advanced Placement (AP) course in statistics (The College Board, 2016), up over 6% from the previous year. Many other students take the course, but do not take the final exam.

In comparison, in 2010 the total enrollment in "calculus-level" mathematics courses in colleges and universities was approximately 886,000, (Blair et al., 2012). Note, though, that this includes up to three semesters of calculus, differential equations, linear algebra, and discrete mathematics, so the actual number of students, as opposed to enrollment, was considerably smaller. Moreover, this total was growing at an annual rate of about 5%. In addition, about 433,000 high school students took the AP calculus final exam in 2016, an increase of about 3% over the previous year according to The College Board, (2016).

Consequently, it appears that there are many more students in the U.S. who are taking introductory statistics courses each year than there are in calculus courses, and the growth rates appear to be higher in statistics. In turn, this suggests that it is becoming increasingly important to focus attention and resources on improving student learning of statistics, a subject that tends to be far more counter-intuitive than calculus and one that tends to attract students who are mathematically less sophisticated than those who take calculus.

Moreover, at most large universities in the U.S., introductory statistics courses are given either by a statistics department with statistically-trained faculty or by individual departments (such as economics, business, biology, psychology, etc) that use statistics heavily, so the faculty are well versed in statistics. At some small colleges, the introductory statistics courses are given by trained statisticians who are often in the mathematics department.

However, at many schools, particularly two-year colleges and most four-year colleges, introductory statistics courses are offered by the mathematics department with instructors (often part-time faculty) having little, or no, formal statistical training. According to the 2010 CBMS study (Blair et al., 2012), over 400,000 students took introductory statistics in a mathematics department; this number represents about four times the number taking such a course through a statistics department. Moreover, the AP (Advanced Placement) Statistics program in the high schools has been growing extremely rapidly (The College Board, 2016); in 2002, about 50,000 students took the AP exam at the end of the course, and this number had been growing at an annual rate of about 25%. As with the college mathematicians who teach the introductory statistics courses, most of the high school teachers have had very little, if any, formal statistical training (though many have had a short teacher development workshop in statistics). It is these two groups to whom this article is primarily directed.

It is not just that these individuals' training has been almost exclusively in mathematics or mathematics education, but that most of their professional focus has been mathematically related as opposed to statistically related. As such, they may be familiar with articles in the mathematics journals, but rarely any in publications focused on statistical education such as the online *Journal of Statistics Education*. As such, they are often not particularly current in terms of statistical education.

3 DYNAMIC PROGRAMS IN EXCEL

Two key notions that underlie virtually every concept and method in statistics are randomness and variation among samples. Has the sample been collected in a truly random fashion that reflects the underlying population? How representative of that population is this one sample? How does this single sample compare to other possible samples drawn from the same population?

Fortunately, most of the critical topics in probability and statistical inference can be dramatically presented using computer graphics simulations to allow students to visualize the underlying statistical populations and so enhance their understanding of the statistical concepts and methods. Many years ago, we addressed the challenge of making these notions evident to students by developing a comprehensive package of computer graphics simulations using BASIC that addressed virtually every topic in introductory probability and statistics. Unfortunately, BASIC routines have become outmoded over the years, especially in terms of being able to provide students (or even colleagues) with copies of the files to explore the statistical concepts on their own computers.

We have lately returned to this challenge and have developed a much more extensive package of dynamic interactive graphical simulations of the mathematics (DIGMath) using Excel because it is available on almost all computers today. Moreover, Excel has the further advantages that it does not require any plug-ins or internet connections, as for example, using Flash, Java, or JavaScript, and it is the standard technology tool for virtually every discipline outside of mathematics. (The complete package can be downloaded from our websites (S.P. Gordon & F.S. Gordon, 2017) or (F. S. Gordon & S.P. Gordon, 2017) for use either by instructors for in-class demonstrations of these ideas or by students for individual or small-group investigations and/or projects, as discussed at the end of this article.)

In the intervening years, many other statistics educators have thought deeply about these issues and have developed specialized software tools to implement random simulations of many of the basic statistical ideas and techniques. These issues have been discussed in a variety of articles, primarily in the statistics literature, as summarized by Mills (2002) for the early years, as well as in individual papers in the *Journal of Statistics Education*. All of the articles in volumes 1 through 24 of the latter from 1993 through 2014 can be accessed from the journal's website, The *Journal of Statistics Education*, (1993-2015); subsequent issues from 2015 (volume 25) on can be found at The *Journal of Statistics Education*, (2016 -). There are many applets available on the Web and many of the same kinds of simulations can be generated using statistical software packages such as Minitab. These efforts have had considerable impact on the teaching of statistics among professional statisticians, but probably much less impact on the teaching of statistics by mathematicians or by statistical users in other disciplines. In part, this may be due to the fact that mathematics faculty without statistical training are often uncomfortable with using such statistical packages in the classroom; others feel it is not appropriate to require students to purchase specialized software in addition to an expensive textbook and calculator; others feel that the one software tool that almost all of their students will someday use in other courses and on the job is Excel (and the majority of students today appear to arrive on campus already knowing how to use it). As a consequence, there seem to be good reasons to have such simulations available in Excel, so that both faculty and students can have easy access to their use without needing to expend money or time in learning new programs.

In this article, we discuss ways in which such dynamic software can be used to enhance student understanding of the fundamental concepts related to regression and correlation. As will be illustrated later in the article, these spreadsheets all use sliders and other Excel *controls* to provide dynamic effects in which parameters and statistical measures can be changed and the resulting effects can be seen virtually instantaneously. Some utilize random simulations to illustrate the variations that occur in the results due to different samples.

In particular, we will discuss the various aspects of regression and correlation that can be enhanced using such software. Many of the same things can certainly be done using the applets available on the Web; however, we focus specifically on ways to use such software as teaching and learning tools to enhance student understanding of the statistical ideas and methods, something that is not always incorporated in the on-line software.

4   SIMULATIONS FOR REGRESSION ANALYSIS

We begin by considering the notion of linear regression. Virtually every problem in statistics textbooks gives the students a single set of data and asks them to construct the regression line based on that data and perhaps use the equation to answer some predictive questions. Regression analysis also arises in all courses in the lab sciences, where students similarly collect one set of sample data and have to construct (possibly only graphically) a line that fits the data. Similarly,

as mathematical modeling becomes ever more prominent in high school algebra, college algebra, and precalculus courses, students are also expected to fit a line (or other non-linear function) to a set of data, assess how good the fit is, and use the function so constructed as a model to answer predictive questions. In all of these settings, most students rarely appreciate the fact that the set of data they have is just one possible set of bivariate data for the two variables under consideration. Rather, they get so involved in finding the line (or curve) that fits the data, either graphically or using technology or, worst yet, by performing the actual calculations, that they lose sight of the underlying statistical ideas.

Furthermore, in the process of discussing the regression line to fit a set of data, particularly laboratory data, it is natural to emphasize that there could be many different sets of data, each leading to a different regression line. A simple computer graphics simulation provides the visual support to make these different lines come to life and provides the means for an investigation on the effects of using different sample sizes. We have developed a very effective version of such a simulation in Excel that is available to any interested reader. It allows the user to select the sample size n (between 2 and 40) and the number of such samples (between 5 and 20). It then generates repeated random samples of size n from an underlying bivariate population that falls into a roughly linear pattern, as shown in Figure 1. The program then calculates the regression line for each random sample and displays the line graphically as illustrated in Figure 2.
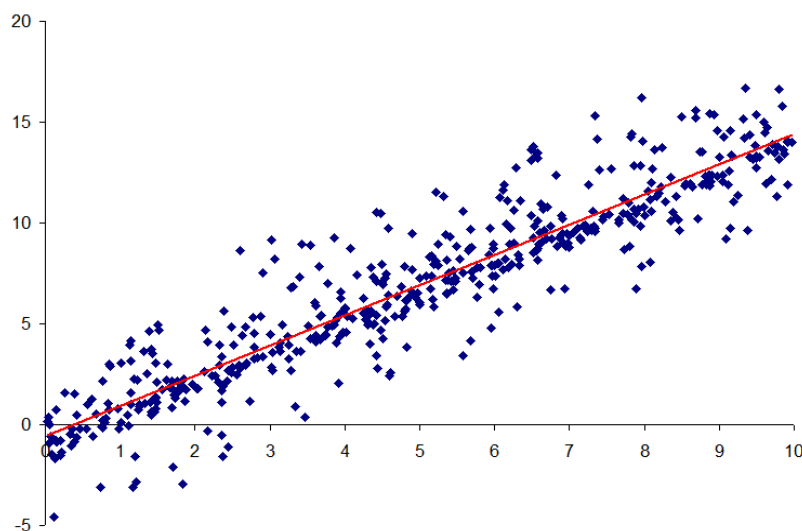.

Figure 1   The underlying bivariate population

For comparison, the simulation also shows the regression line for the underlying population. In addition, the point (μ_x, μ_y) corresponding to the means μ_x and μ_y of the x-values and the y-values, respectively, is highlighted by the circle in the center. Recall that the regression line always passes through the point $(\overline{x}, \overline{y})$ corresponding to the mean of the x's and the mean of the y's, so that this highlighted point certainly lies on the population regression line.

In Figure 2, we show one set of results from this dynamic simulation that draws 10 random samples of size $n = 3$. Figure 3 shows a different run of the simulation, also with 10 samples of size $n = 3$. Obviously, the results are dramatically different.
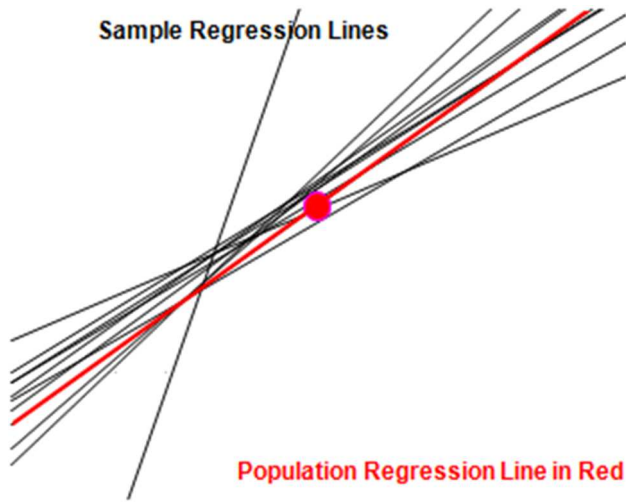
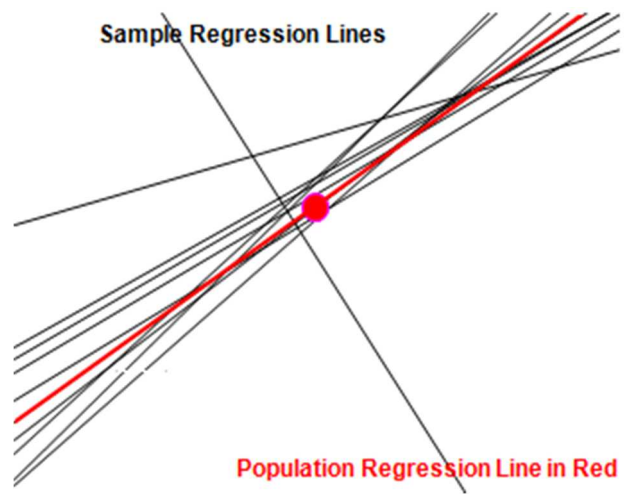Figure 2 Simulated regression lines based on samples of
$n = 3$ points

Figure 3 Simulated regression lines based on samples of
$n = 3$ points

Notice that the majority of the sample regression lines tend to be more or less similar to one another in terms of overall slope and position. However, there are some regression lines whose slopes vary dramatically from that of the population's regression line (the heavy line in the figure) and so are at very sharp angles to the population line; in fact, some of them are almost perpendicular to that line. With samples of size $n = 3$, one can easily visualize selecting three points whose x-values are reasonably close to one another with relatively large differences in the y-values, so that the corresponding regression line would be at a steep angle compared to the population regression line. However, this is very unlikely to happen with larger sample sizes, so that one should expect that the sample regression lines based on larger samples should be more closely aligned to the population regression line. This image conveys the important message that regression analysis based on small samples can be highly suspect. From a pedagogical standpoint, this kind of display can be very informative in helping the students see that each sample gives rise to a different regression line, though most of them remain relatively close to the population regression line, (at least when the sample size n is small), and most of them pass fairly close to the indicated point (μ_x, μ_y). Furthermore, whenever a new set of samples is run (seemingly instantaneously at the push of a button), the resulting image will change dramatically as very different samples are likely drawn and hence very different sample regression lines are generated.

When we use larger sample sizes, almost all of the sample regression lines have slopes that are very close to that of the population regression line and so the resulting sample regression lines usually lie much closer to the population line, as shown in Figure 4 with $n = 25$ points. In turn, this reinforces the notion that increased sample size makes for far better predictions. Moreover, the fan-shaped pattern seen here is quite typical of what happens – most of the lines have very similar slopes and most pass very close to the point (μ_x, μ_y). However, as one moves further away from this point, the various lines certainly tend to diverge from one another.

Moreover, particularly with larger sample sizes, it becomes evident that almost all of the sample regression lines pass very close to the point (μ_x, μ_y). For each sample, we expect that the mean, $\overline{x}$, of the x's should be a good estimate for the mean μ_x of all the x-values in the population and that the mean, $\overline{y}$, of the y's should be a good estimate for μ_y. As a result, we should expect that, for any given sample regression line, the corresponding point ($\overline{x}$, $\overline{y}$) on it should be quite close to the point (μ_x, μ_y) on the population's regression line. In turn, this explains why almost all of the sample regression lines pass close to the same point.
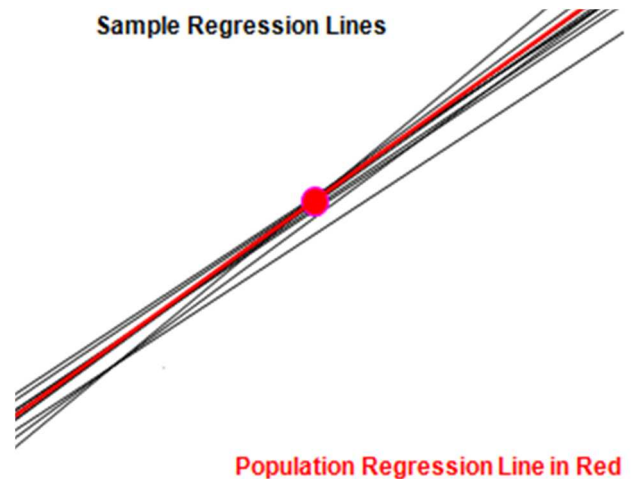

Figure 4 Samples regression lines with $n = 25$

This kind of investigation also provides a wonderful opportunity to stress two of the key themes in statistics – the effect of sample size on the outcome and the variation that occurs both within a sample and between different samples.

www.manaraa.com

5 SIMULATING THE CORRELATION COEFFICIENT

We next turn to a consideration of the correlation coefficient r, which is used as a measure of how well a line fits a set of data. In particular, the possible values for r range from a minimum of -1 to a maximum of +1. Values of r close to 1 indicate a high level of positive correlation between the variables (the slope of the regression line is positive and the points are tightly clustered about the line) and values of r close to -1 indicate a high level of negative correlation (the slope of the regression line is negative and the points are also tightly clustered about the line). Values of r close to 0 indicate no correlation and the points are widely distributed without any particular pattern about the regression line.

Furthermore, the square of the correlation coefficient, r^2, the coefficient of determination, indicates the percentage of the vertical variation in the data about the line that can be attributable to the linear model. Thus, if $r = 0.8$, then $r^2 = 0.64$ and hence 64% of the variation can be explained by the linear function, so that 36% is not so explained.

In order to determine whether the value r for the correlation coefficient based on a given set of data is "close enough to +1" or "close enough to -1" to conclude that there is indeed a high level of correlation between the two variables, one compares the value of r obtained to an appropriate critical value for the correlation coefficient from a statistical table. Each of the critical values corresponds to a different sample size n and, the larger that n is, the smaller the critical value. Thus, for a sample of size $n = 5$, the critical value is 0.878, so that the correlation coefficient for the data must be larger than this number in order to conclude that there is a statistically significant level of correlation. Similarly, for a sample of size $n = 15$, the critical value is 0.514, so it is a much lower bar that one must get over to conclude statistical significance. The reason is that the larger sample size provides considerably

more information and hence one needs less conclusive evidence. Honestly, these are fairly sophisticated notions that likely go over the heads of many students in introductory statistics.

To provide deeper insight into these ideas, we have developed another dynamic spreadsheet to simulate the results of examining the correlation coefficients associated with many different bivariate samples drawn from the same underlying population used in the regression simulation. The user has the choice of the sample size n, between 3 and 50, and the number of random samples, between 50 and 250. The program then generates repeated random samples of that size, calculates the correlation coefficient for each random sample, and displays the values graphically in a bar chart to show their distribution. The program also displays the mean and standard deviation of the correlation coefficients from the random samples and allows the user to compare that mean to the correlation coefficient of the underlying population.

In Figure 5, we show the results of one run of this program with 200 samples of size $n = 7$. Notice that the resulting set of sample correlation coefficients seem to be widely scattered across the histogram from about $r = 0.7$ to $r = 1$ with the bulk of them clustered toward the right side with values above $r = 0.9$. (Occasionally, there will be a small cluster at the far left of the bar chart, which indicates samples for which there is little or no correlation; these are samples for which r is 0.5 or smaller.) Moreover, the correlation coefficient for the underlying population is $r = 0.926$; for the samples in Figure 5, the mean of the 200 correlation coefficients was 0.919 with a standard deviation of 0.082. Despite the wide variation in the values of the sample correlation coefficients, their average is surprisingly close to that of the population and the standard deviation is fairly small. Additional runs of the programs produce very comparable results, though the actual distribution will change fairly dramatically.
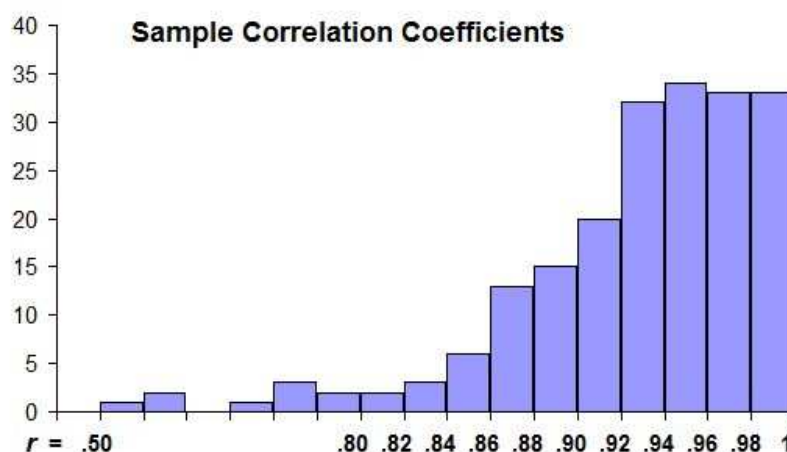

Figure 5 The correlation coefficients of 200 random samples of size $n = 7$

In Figure 6, we show the results with 200 samples of size $n = 20$. Notice that almost all of these sample correlation coefficient values are clustered fairly closely about the population correlation coefficient value $r = 0.926$ in a shape that begins to look somewhat normal (although it is definitely skewed to the left). This means that most of the regression lines associated with these samples have fairly comparable slopes. The mean of the 200 correlation coefficients is 0.924, which is quite close to the population value $r = 0.926$, and the associated standard deviation is 0.036, which reinforces the conclusion that the distribution is much more densely packed about the population value. Also, repeated sets of samples of the same size have the same properties.
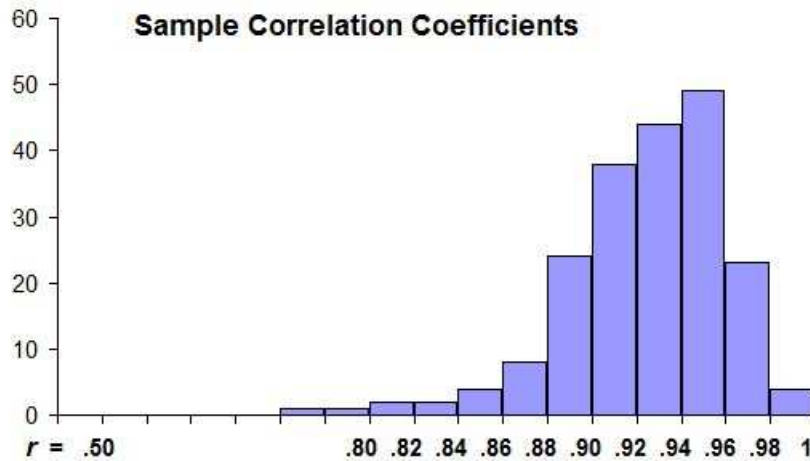

Figure 6 The correlation coefficients of 200 random samples of size $n = 20$

Furthermore, Figure 7 shows the results of 200 samples of size $n = 50$, where we see that the correlation coefficient values are much more densely packed and the shape of the bar chart is even more suggestive of a normal shape. The corresponding mean of these 200 correlation coefficients is 0.928 with a standard deviation of 0.018. Therefore, we see that, as the sample size increases, it becomes evident that the resulting sample correlation coefficients become ever more tightly clustered about the population correlation coefficient.
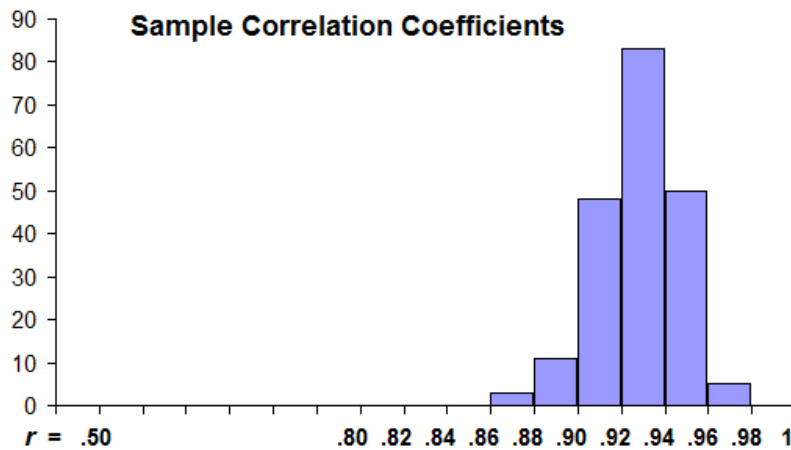

Figure 7 The correlation coefficients of 200 random samples of size $n = 50$

Moreover, as the sample size n gets closer to the maximum value of 50 allowed in this program, notice that the shape of the distribution of correlation coefficients looks more and more normal. But it is always skewed because there is a fixed cut-off at the right, namely $r = 1$ and the underlying population has a correlation coefficient that is quite close to 1, while there can be a tail to the left. Moreover, the mean of the sample values generated in the program are typically very close to the value of r for the underlying population, for any value of the sample size n. Finally, notice how the value for the standard deviation typically gets smaller as the sample size increases. This simply reflects the fact that, for larger samples, the values of the correlation coefficient are clustered ever more tightly about the population value and the spread diminishes dramatically. Note, though, that merely increasing the sample size does not necessarily mean that the average of the sample correlation coefficient values will automatically be closer to the population correlation coefficient nor does it mean that the standard deviation of those sample correlation

www.manaraa.com

coefficient values will automatically be smaller. This is an important point to emphasize to students.

6   EFFECTS OF AN EXTRA POINT ON THE REGRESSION LINE AND THE CORRELATION COEFFICIENT

We next consider the effect that an extra point, say an outlier, can have on both the regression equation and the correlation coefficient using a third dynamic program. In this

module, the user can experiment with an underlying set of 5 points, 10 points, and 20 points, each found on a separate page (via a different tab at the bottom) of the spreadsheet. The program draws the scatterplot of those data points and the associated regression line, along with the equation of that line and the corresponding correlation coefficient. The location of the "extra" point is controlled by a pair of sliders that allow the user to move the point horizontally and vertically. As it moves, the associated regression line based on the extended set of points is drawn and the corresponding equation of the line and the correlation coefficient are displayed, as illustrated in Figure 8.
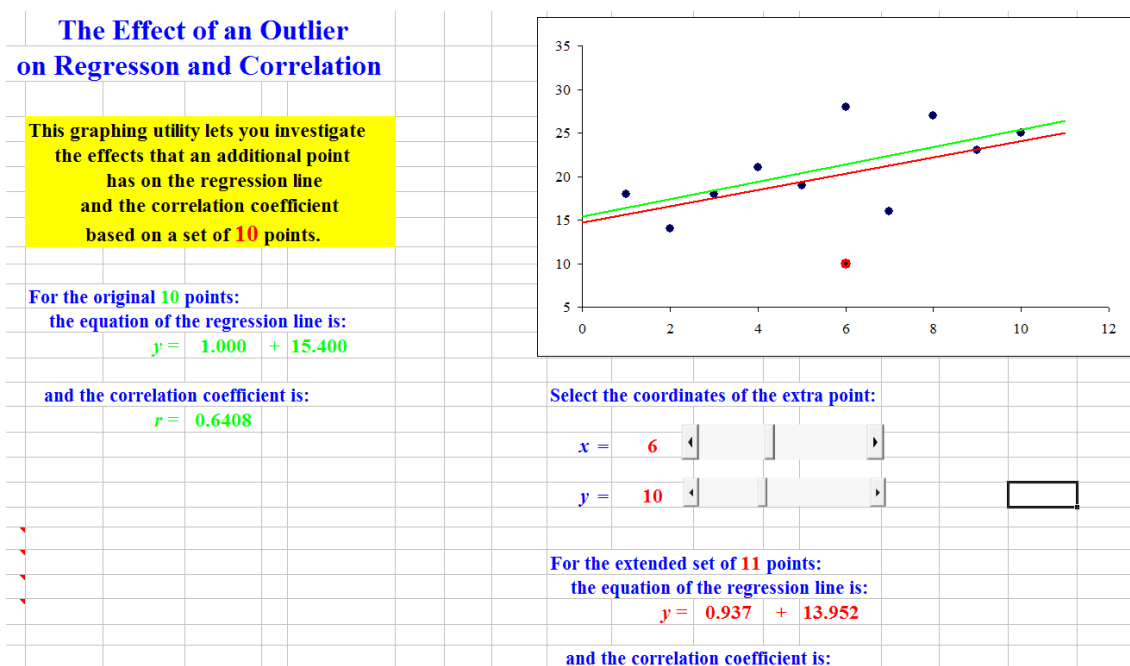

Figure 8 The effects of an extra point

Students can easily see the effects of the position of the extra point. When it is positioned relatively close to the regression line for the underlying set of data, the two regression lines are, visually, very close and the coefficients in both equations are numerically very close to one another. Also, the values for the correlation coefficients are close to one another. As the extra point moves farther and farther away from the underlying regression line, the differences become more and more pronounced:

- the slopes of the two regression lines become significantly different and hence the lines diverge more and more sharply from one another;
- the intercepts of those lines become significantly different;
- the two equations differ more and more from each other; and
- the values for the two correlation coefficients become more and more different.

Moreover, when comparing the same kinds of changes when there are 5, 10, and 20 underlying points, students see

that the effects of the extra point are greatly magnified when there are only five points and are less and less pronounced as the number of points increases. This reinforces the notion that effects of an outlier are diluted with larger sample sizes, a fundamental concept that definitely needs repeated emphasis in different contexts for many students.

7   VISUALIZING THE SUM OF THE SQUARES

We next consider the use of the sum of the squares as an alternate way of measuring the goodness of the linear fit, so that students have a tool other than the correlation coefficient with which to make a decision. Because there is a simple geometric image associated with it, most students in introductory statistics and in precalculus and algebra courses that incorporate curve fitting as a major topic find it more meaningful than the correlation coefficient that is calculated by some rather arcane formula. In turn, when faced with the problem of finding the parameters – either the slope and intercept of a line or the parameters associated with some

nonlinear function – to improve on the fit of a function to the data by reducing the sum of the squares value, many students react to it as a challenge and excitedly experiment with the effects of varying the parameters to produce a better fit than anyone else in the class.

As an exploratory tool to investigate the sum of the squares, one of the authors has developed still another dynamic spreadsheet that visually displays the sum of the squares associated with a set of data. The user can select between 2 and 20 data points and then has to enter the

coordinates for those points. The program draws the scatterplot. The program asks the user to supply values for the slope and the vertical intercept of a possible line that fits the data and the corresponding line is then superimposed over the scatterplot. In addition, the program displays the value of the sum of the squares. See Figure 9. This allows students to experiment with changing the equation of the line via its parameters and seeing both visually and via the value for the sum of the squares how well the line comes to fitting all the data points.
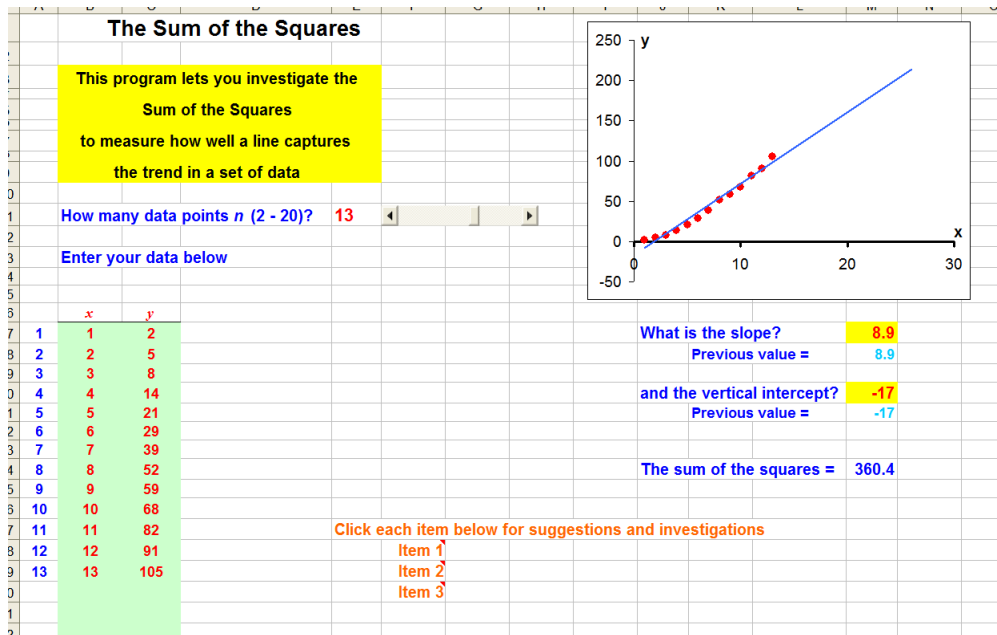

Figure 9 The sum of the squares

8   OTHER LINES THAT FIT THE DATA

In addition to the topics discussed above, the authors have also developed comparable dynamic spreadsheets to allow students and/or instructors to experiment visually with a variety of other topics dealing with the notion of fitting a line to a set of data. One of these is a graphical implementation of the median-median line, which is implemented on many graphing calculators; this is based on creating a line by:
1   partitioning the data into three groups (the lowest, the middle, and the greatest),
2   finding the median of each group,
3   using the two points corresponding to the first and the third median values; and
4   adjusting that line based on the second median value to produce a line to fit the data.

A second spreadsheet is a random simulation of the median-median line that generates many random samples from the underlying bivariate population, calculates and draws the corresponding median-median lines, and displays the associated statistical values.

In a similar way, a third program simulates the quartile-quartile line, which creates a line to fit a set of data by finding the first and fourth quartiles of the data and finding the line that passes through those two points.

A fourth program is a simulation of all three best-fit lines in which the user can select the sample size (between 4 and 40) and the program generates a random sample of that size from the same underlying bivariate population used previously and displays the associated scatterplot. The program then calculates and displays the regression line, the median-median line, and the quartile-quartile line, as well as the value for R^2, the coefficient of determination for each of the three lines to provide a measure of how well each line fits the sample data. The idea is to compare how well each of the three lines, each based on a different approach, fits the data. While we do not go into any further details here, all of these programs are available for download from the authors' websites.

9 RELATED CURVE FITTING ROUTINES

One of the present authors (S.P. Gordon, 2017) has also created a variety of related dynamic spreadsheets for demonstrating and experimenting with nonlinear curve fitting. One fits the best exponential function to a set of data entered by the user; another fits the best power function to a set of data. A third generates the best linear, exponential, and power function to fit a set of data, so that the user can see and compare all three to decide which seems to be the best choice among the different candidates. Still another performs polynomial regression where the user has the choice of the desired degree. And, finally, one fits a sinusoidal function (either a sine or a cosine) to a set of roughly periodic data.

10 CONCLUSIONS

For most students in introductory statistics, the fundamental concepts regarding regression and correlation are quite sophisticated, particularly the ways in which the result can vary dramatically from one sample to another. In addition, the manner in which the results change as a function of the sample size can also be challenging for many of these students to graph. The use of dynamic software simulations (whether the ones developed by the authors or others available on-line) can provide invaluable teaching and learning tools to increase student understanding of the ideas and help them master the concepts in ways that make the techniques more understandable and hence adaptable to many different situations that will arise in other courses and in increasingly many careers.

All of the dynamic spreadsheets mentioned here, as well as many others covering virtually every topic in introductory statistics and probability, can be downloaded from the authors' websites. If interested readers have any suggestions for additional topics they would like to see treated, they are encouraged to contact the authors with their suggestions.

REFERENCES

College Board, (2016) *AP Program Participation and Performance Data 2016,* The College Board, URL: https://research.collegeboard.org/programs/ap/data/particip ation/ap-2016

Blair, Richelle, Ellen E. Kirkman, and James W. Maxwell, Eds., (2012) *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States, The Fall 2010 CBMS Survey*, American Mathematical Society, Providence, RI. ISBN 978-0-8218-9412-5

Gordon, Sheldon P. and Florence S. Gordon, (November, 2017) Dynamic spreadsheets for statistics and probability, Farmingdale State College of SUNY, URL: http://www.farmingdale.edu/faculty/sheldon-gordon/dynamicstatistics.shtml

Gordon, Florence S. and Sheldon P Gordon, (November, 2017) Dynamic spreadsheets for statistics and probability, New York Institute of Technology, URL: http://iris.nyit.edu/~fgordon/StatisticalSimulations.htm

Gordon, Sheldon P., Dynamic spreadsheets for college algebra and precalculus, (September, 2017) Farmingdale State College of SUNY, URL: http://www.farmingdale.edu/faculty/sheldon-gordon/dynamicprecalculus.shtml

Journal of Statistics Education, (1993-2015), **1-24**, American Statistical Association, URL: http://ww2.amstat.org/publications/jse/jse_archive.htm

Journal of Statistics Education, (2016 -), **25-** , Taylor & Francis, URL: http://tandfonline.com/toc/ujse20/current

Mills, Jamie D., (2002) Using Computer Simulation Methods to Teach Statistics: A Review of the Literature, *Journal of Statistics Education*, **10**

BIOGRAPHICAL NOTES

Sheldon Gordon is SUNY Distinguished Teaching Professor of Mathematics at Farmingdale State College in New York, USA. Throughout his career, he has been deeply involved in efforts to revitalize the undergraduate mathematics curriculum, particularly from the perspective of the use of technology to enhance student understanding of mathematical concepts and methods. Shelly has authored more than 200 papers, a dozen textbooks and monograph volumes, and hundreds of software programs for mathematics education.

Florence S. Gordon is a retired Professor of Mathematics at New York Institute of Technology in New York, USA. She is a co-author of *Functioning in the Real World*, co-author of *Contemporary Statistics: A Computer Approach* and co-editor of two MAA Notes volumes, *Statistics for the Twenty First Century* and *A Fresh Start for Collegiate Mathematics: Rethinking the Courses Below Calculus*. She has published extensively in mathematics and statistics education.