TEACHERS' KNOWLEDGE OF STUDENTS' CONCEPTIONS AND THEIR DEVELOPMENT WHEN LEARNING LINEAR REGRESSION

Stephanie A. Casey
Eastern Michigan University, Ypsilanti, Michigan USA
scasey1@emich.edu

This paper synthesizes the results of three research studies centered around the teaching and learning of linear regression. In addition to regression's practical importance, in many cases linear regression is students' first experience with the fundamental concepts of statistical association and dependence. Two of the studies researched how students progress in their learning of linear regression, from the initial conceptions of eighth grade students to the understanding expected of twelfth grade students. The third study examined the work of teaching grade nine students informal line of best fit, their first introduction to linear regression, to generate a description of the pedagogical content knowledge needed by teachers to effectively guide students towards understanding of the topic.

THE STUDIES

Statistical association between two variables, also known as covariation, is one of the fundamental statistical ideas in school curricula (Burrill & Biehler, 2011; Garfield & Ben-Zvi, 2004). In fact, reasoning about statistical association is one of the most important cognitive activities that humans perform (McKenzie & Middlesen, 2007). Therefore, it is crucial that teachers have the statistical knowledge for teaching (Groth, 2013) statistical association effectively. Necessary components of statistical knowledge for teaching include knowledge of students' conceptions and how these conceptions can develop into robust understanding of the topic with proper learning experiences designed and implemented by teachers. This paper synthesizes the results of three research studies that focused upon this component of statistical knowledge for teaching linear regression, the primary topic students study to learn about statistical association of quantitative data. This section provides descriptions of the three studies.

A hypothetical learning trajectory (HLT) for teaching and learning linear regression at the middle and secondary school levels was a product of the research study Project-SET[1] (see Bargagliotti et al. (2012) for the complete HLT). A HLT provides a model for the successive and gradual thinking a learner must go through to achieve deep understanding of a topic (Duschl, Schweingruber, & Shouse, 2007). This HLT was created through an iterative process of interaction between a set of practitioners and the research team, of which I was a member. It was built from research literature regarding how people learn regression as well as observations of those who had taught linear regression. It has a loop structure, depicting a spiraling process of learning linear regression by progressing through 5 loops. Each loop builds upon the learning that has occurred in the previous loop. Each loop has a specific learning focus: loop 1 focuses on informal line of best fit; loop 2 is about the ordinary least squares regression line (hereafter referred to as regression line); correlation is the focus of loop 3; regression equations from samples are studied in loop 4; and loop 5 focuses on the sampling distribution of the slope of the sample regression line. The loops will be described in greater depth in the body of the paper. The HLT maintains a consistent organization for the learning sequence in each loop by having the learner proceed through the statistical investigatory process (Franklin et al., 2007). Thus, as learners progress through each of the five loops, they are formulating a statistical question appropriate for that loop's focus, collecting data, analyzing the data, and interpreting the results of their analysis.

The first loop of the HLT initiates the learning of linear regression through the study of the informal line of best fit. The informal line of best fit is included in many current national curriculum standards (e.g., Australia: Australian Curriculum, Assessment, and Reporting Authority, 2012; England: Qualifications and Curriculum Authority, 2007; U.S.A.: National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) as a natural extension of students' developing knowledge of linear equations and an introduction to statistical association. For example, the authors of the Common Core State Standards for Mathematics (CCSS-M) (National Governors Association Center for Best Practices & Council of

Chief State School Officers, 2010) state that Grade 8 students should “know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, [students will] informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line” (p. 56). The widespread inclusion of informal line of best fit in school curricula as well as its important role in beginning students’ study of statistical association led me to direct two research studies concerning the teaching and learning of informal line of best fit[2]. One study involved videotaped task-based interviews with 33 Grade 8 students to understand their conceptions of the meaning of line of best fit, its criteria, methods for finding it, as well as sources of these conceptions. The interviews were done in the week preceding the students’ formal instruction on informal line of best fit in their mathematics class, so the students had completed their preparation for the learning of the topic prior to the interviews. Each interview began by presenting five tasks that asked the participating student to use pieces of wire to mark the informal line of best fit for contextual data presented in a scatterplot. Each student was asked to talk aloud during the process, and was also asked follow-up questions regarding criteria and methods used in the process. The five plots varied in their direction of association as well as scatter of the points. The sixth task asked each student to decide which of two presented lines (see Figure 1) is a better fit for the given data and to explain the reason for that choice. Line 1 was positioned generally near the points and specifically through two of them while Line 2 was placed closest to all of the points but not through any of them.

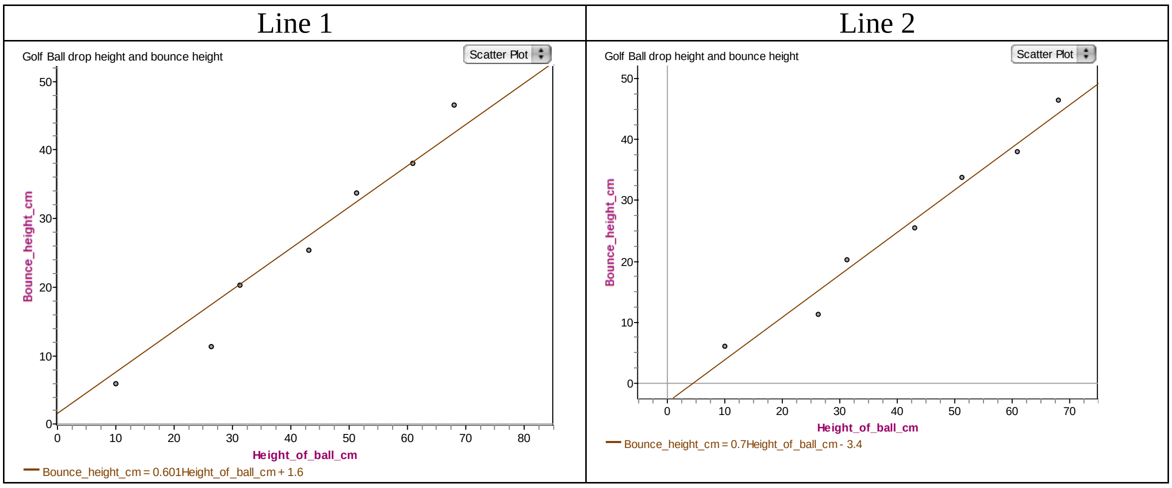

Figure 1. Student interview task six: Which line fits the data better and why?

The interview also included six questions which asked participants to reflect upon their responses and define the line of best fit. The data were studied by two analysts through an iterative process, moving back and forth between viewing of the videotapes and identifying, discussing, and classifying the responses, which allowed themes to emerge.

The purpose of the other study was to generate a detailed and comprehensive description of the statistical knowledge for teaching informal line of best fit. This paper presents relevant results regarding pedagogical content knowledge only. The methodology used for this study involved the generation of data at both a primary and secondary level. The primary level documented the work of teaching informal line of best fit through a case study of a secondary mathematics teacher as she taught the topic to grade nine students over five class sessions. Primary level data included transcripts of the observed class sessions, transcripts of interviews done with the teacher immediately following each class session, copies of the relevant textbook and teacher manual pages, and assessment materials. A team of analysts, including statisticians and statistics educators, used the primary level data as a catalyst for developing conjectures regarding the knowledge teachers need to effectively teach informal line of best fit. The analysts’ conjectures regarding the needed knowledge comprise the data generated at the secondary level.

The rest of this paper will synthesize the results from all three studies (Project-SET HLT, student conceptions, and statistical knowledge for teaching informal line of best fit) regarding teacher knowledge of students' conceptions and their development when teaching linear regression, including informal line of best fit, regression line, correlation, and sampling distributions for the slope of regression lines from a population.

INFORMAL LINE OF BEST FIT

Prior to studying linear regression, the HLT expects that learners have developed intuition about bivariate data through approaches like stacking and color gradients (see Konold & Higgins (2003) for examples), and can determine the slope, y-intercept, and equation of a line. Learners' study of linear regression begins in loop 1 of the HLT with the formulation of a statistical question that explores a relationship between two quantitative variables measured on each unit in a group of interest. Next, learners determine a method for obtaining the data to answer the question and carry out the data collection. Since informally determining the line of best fit is a visual process, one needs to convert data recorded in tabular form into a graphical scatterplot of response against predictor. Teachers need to know difficulties students have with this transition from a tabular representation to a graphical representation. First, students have difficulty identifying which variable is the predictor variable, and they also need experience with examples in which the classifications are ambiguous to understand that this may be the case in some instances. Students may also have problems when creating a scale for the axes. In particular, students may initially create scales that are not uniform (particularly if the data themselves are not uniformly distributed across the range) or do not capture all the data in a meaningful way. For example, students in the case study class of the teacher knowledge study were asked to graph bivariate data concerning the amount of supplies on days 2, 5, and 9, then asked to predict the amount of supplies on day 15. Some students spaced the values of 2, 5, and 9 equally, making the gap between days 2 and 5 appear the same size as the gap between days 5 and 9. Also, many students needed to be reminded by the teacher that the scale should enable the prediction of supplies on day 15.

In helping students develop their abilities to read a scatterplot in order to visually determine and describe relationships between the two variables, teachers need to know that students need assistance navigating differences they find confusing between plotting functions and plotting data. For example, in a function no x-value will have more than one corresponding y-value, but in a data set this is possible. Students in the class observed during the case study found this confusing. Some advocated for averaging the values of the y-variable if this happened in order to plot a single point at that x-value. Another difference is that linear functions are either monotonically increasing or decreasing, but this is not necessarily the case for data with a linear association, nor does it imply that the data do not have a linear association as some students are inclined to believe. Teachers should also know that students likely will have difficulty seeing overall trends in the data when reading scatterplots, stemming from the fact that novice analysts tend to focus their attention on individual cases in the data and perceive data as a series of individual cases (case-oriented view) rather than considering the entire set of points in the data holistically with characteristics that are invisible in any of the individual cases (aggregate view) as more expert analysts do (Bakker, 2004). The student conception study showed that the case-oriented view was prevalent in younger students who are prepared to study regression for the first time. Many students in this study gave all of their attention to select points, such as the first, last, highest, or lowest points, or a subset of the points, such as two points with the same y-coordinate, and disregarded the rest of the points when reading the scatterplot for trends.

For data with a linear trend, curriculum standards (e.g., Australia: Australian Curriculum, Assessment, and Reporting Authority, 2012; England: Qualifications and Curriculum Authority, 2007; U.S.A.: National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and the HLT call for learners to begin their study of linear regression through learning about informal lines of best fit. The student conceptions study found that students' definitions of the line of best fit fell into four categories: where you expect the relationship between the variables to be; shows what the data looks like; average; and something you use to get close predictions. Those students who conceived of the line of best fit as showing where you expect the relationship between the variables to be had a view that the line represented a general relationship between the two variables of interest, while students with the second conception viewed the line as a way to guess what the data would be if you took away the scatterplot of the actual data points and therefore had more of a localized understanding of the relationship. Other students defined it in terms of the average, using a phrase like "average of the data points". The last category of definition of the best fit line by students in the conceptions study emphasized the use of the line for making predictions. Students with this definition said things like "if you used it for prediction, it would be close". Knowledge of students' conceptions of the line of best fit is needed so that teachers can anticipate what ideas students typically come to the learning of this topic with and plan their instruction accordingly.

Continuing in loop 1 of the HLT, learners devise methods to fit an informal line of best fit to data that indicates a linear relationship. The opportunity to create one's own method is a particularly salient situation where teachers' knowledge of students' conceptions can enable them to manage the real-time demands of teaching this topic. The following summary of the predominant methods and their justifications that emerged from the student conceptions study informs this type of knowledge.

When asked to find the line of best fit, a sizeable number of students in the student conceptions study wanted to bend the wire or connect the points on the scatterplot. Many students struggle to conceive of the line of best fit as a line that does not necessarily go through all of the points, likely because this differs from graphs of linear functions that these students have been studying in mathematics. This relates to the predominant method students used to place the informal line of best fit: through as many points as possible. For some students this was related to their conception of the line of best fit as something you use to get close predictions. Hence, these students thought if you put the line through the most points you will get the most accurate predictions. Students using this method also analyzed the scatterplot by searching for collinear points, rather than viewing the general trend of all of the points. A subset of these students required their lines to start at the origin, with rationale including "that is where all lines start" as well as knowledge of the context for a particular plot (e.g., dropping a ball from a height of zero centimeters should result in it bouncing zero centimeters).

Getting an equal number of points on both sides of the line was the second most common method. This method was common for those students who viewed the line as an average or where you expect the relationship between the variables to be. It corresponds to the calculation of the median, a univariate measure of center these students knew. Interestingly, three students always placed their lines horizontally or vertically so that there were an equal number of points on either side of the line. These students decided to focus their attention on only one of the variables of interest in the scatterplot, and placed the line of best fit to represent the middle of that univariate data set. These students were unable to coordinate the relationship between the two variables in the bivariate setting of statistical association, and thus reduced it to a univariate setting.

Seven of the 33 students used the criteria encouraged by the authors of the CCSS-M (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and in agreement with the approach of the ordinary least squares regression line: as close to all of the points as possible. Students with this approach justified their methods by saying things like the line was placed to be "relatively close to all of the points" or "Really I was trying to make that [the line] go in between so they [the points] are all close". Task six (see Figure 1) of the interview was designed to see whether students viewed it more important for the line to go through some of the points (Line 1) or to be closest to all of the points but not necessarily go through any of them (Line 2). One-third of the students preferred the line that went through some of the points (Line 1) to the line closest to all of the points (Line 2). Thus, teachers can anticipate that a sizeable number of their students will likely need learning experiences to change their conception that it is more important to go through points than to be near all of the points. This conception needs to be addressed in order to enable all learners to meaningfully explore the concept of correlation in loop 1 of the HLT through examining how closely the points follow a line. Finally, other methods used by multiple students included connecting the first and last points, starting from the first dot then maximizing the number of points the line goes through, and knowledge of the context the variables of interest are describing.

The lack of one correct line of best fit when determined informally and the variability associated with the lines created by the students in a classroom can be bewildering to students who are used to working with linear functions where there is one unique line that can be drawn through the set of plotted points. Teachers need to know that students will struggle with variability in the answers to questions posed when beginning their study of linear regression, not only with the variability in line placement but also the variability in their predictions made using their lines. Learners in this loop learn about extrapolations, which many are uncomfortable with because they think it is invalid to make predictions for values outside of the original domain of the data set. Learners need instructor guidance to understand the value but limitation of extrapolations.

To complete loop 1 of the HLT, and in agreement with curriculum standards such as the CCSS-M (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), learners interpret the slope and y-intercept of their informal line of best fit in the context of the posed question. Students' previous study of lines and slope with mathematical functions can create cognitive obstacles to correct interpretation of the slope and y-intercept of best fit lines in a statistical setting. For example, in mathematics it is emphasized that lines have a constant rate of change (slope) that is the amount the y-variable changes by every time the x-variable increases by one unit. This is not the case for the regression line's slope, as it represents the *average* change expected in the predicted values of the response variable for a one unit increase in the predictor variable, rather than the amount the predictor variable necessarily changes by in the deterministic mathematical definition. Also, students need assistance transitioning from a "rise over run" understanding of slope emphasized in mathematics to a "predicted change in y for a one unit change in x" understanding in order to connect the meaning of the slope of a line of best fit to the context. Regarding the y-intercept of the line of best fit, students need assistance understanding that the primary criteria of getting the line of best fit as close to all of the points as possible may necessitate the marking of a line whose y-intercept is meaningless in the context of the problem.

REGRESSION LINE

The HLT describes the developmental progression of learning that builds on the learning of the informal line of best fit in loop 1 to support the understanding of a formal regression line and related topics in loops 2 through 5. Understanding how learning of linear regression develops is a fundamental component of statistical knowledge for teaching linear regression (Groth, 2013). Transitioning from loop 1 to loop 2, learners are expected to use technology to find the regression line, focus their attention on residuals and the residual plot as a way of assessing the accuracy of a line of best fit, and understand how the regression line can be viewed as the one that best fits the data because it minimizes the sum of the squared residuals.

In loop 2, learners are also introduced to calculations to measure variability in the residuals ($\sum (y-\hat{y})^2$) and the values of y ($\sum (y-\overline{y})^2$) and learn how their relative sizes relate to the strength of the association. This forms the foundation for the formal study of correlation in the HLT's loop 3. Attention in loop 3 is centered upon understanding the correlation coefficient, *r*, and the coefficient of determination, r^2, including their calculations and interpretations in the context of the dataset. Emphasis is also placed upon developing learners' understandings of the distinctions between correlation and causation.

The final two loops of the HLT, 4 and 5, build the foundational knowledge learners need to understand inferential procedures for the slope of the regression line. Loop 4 changes the statistical question of interest from a scenario with access to the entire population to one where you cannot obtain the population data but can take a random sample from the population. In this new scenario, one is expected to determine an appropriate sampling method and use it to obtain a sample from the population. Next, learners use technology to find the regression line for the sample data and learn that the sample regression equation's slope and initial value are estimates of these corresponding parameters in the population's regression model. Finally, in loop 5, learners build on their understanding of linear regression and concepts of inference to study the slope of regression lines created by repeatedly sampling the same population. This requires another level of abstraction and sophistication, as learners are asked to study the collection of slopes of regression lines as another object in and of itself. Learners encapsulate this understanding by viewing and studying this collection as the sampling distribution for the slope of the sample regression line.

CONCLUSION

Taken together, the results of these three studies present a detailed description of the knowledge needed by those who teach linear regression regarding learners' conceptions of the topic and how they develop. This is the fundamental component of teachers' pedagogical content knowledge for teaching linear regression, for it provides the basis for the remaining components of knowledge of curriculum and knowledge of content and teaching (Groth, 2013). The description makes clear that the depth of knowledge concerning student learning needed by teachers who teach linear regression is substantial and differs in significant ways from the knowledge needed to teach linear functions. This speaks to the larger need for school teachers of mathematics, who are responsible for teaching statistics since that is where it is included in school curricula, to be distinctly and thoroughly prepared for teaching statistics.

NOTES

1. This work is supported by National Science Foundation Grant No. 1119016.
2. Members of the research team whose work contributed to the findings described in this paper are David Wilson, Jennifer Kaplan, and Adam Molnar.

REFERENCES

Australian Curriculum, Assessment, and Reporting Authority (2012). *The Australian curriculum: Mathematics.* Sydney, Australia: Author.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83.

Bargagliotti, A., Anderson, C., Casey, S., Everson, M., Franklin, C., Gould, R., Groth, R., Haddock, J., & Watkins, A. (2012). *Project-SET linear regression learning trajectory*. http://project-set.com/presentations/121712-regressionlp-final-released/

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In Batanero, C., Burrill, G., & Reading, C. (Eds.), *Teaching statistics in school mathematics - Challenges for teaching and teacher education. A Joint ICMI/IASE study: The 18th ICMI study* (pp. 57–69). New York, NY: Springer.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: National Academy Press.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Schaeffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A PreK-12 curriculum framework.* Alexandria, VA: American Statistical Association.

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 397–409). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Groth, R. E. (2013). Characterizing key developmental understandings and pedagogically powerful ideas within a statistical knowledge for teaching framework. *Mathematical Thinking and Learning, 15*(2), 121–145.

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A Research Companion to Principles & Standards for School Mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54,* 33–61.

National Governors Association Center for Best Practices & Council of Chief State School Officers (2010). *Common Core State Standards (Mathematics).* Washington, DC: Authors.

Qualifications and Curriculum Authority (2007). *The National Curriculum 2007.* Coventry, England: Author.