Project-SET Linear Regression Learning Trajectory June 2014 [1,2,5]

	GAISE Framework[4]				
	Formulate Question	**Collect Data**	**Analyze Data**	**Interpret Results**	**Key Developmental Understandings**
Loop 1	How can we display a bivariate quantitative distribution and extend the concepts of shape, center, and variability to summarize it?	a. Determine a method of obtaining the measurements/counts for each variable from each member of the group	a. Plot the data in a scatterplot[6] b. Discuss reasons why there is "scatter" in the plot c. Visually determine and describe the trend of the plot d. Discuss ways to model the data e. Examine scatterplots and make attempts to draw a model on the graph (e.g., this could be done by drawing a line, or a curve on the scatterplot) f. For those data where a linear model is appropriate, discuss where the "line of best fit" would be placed[6]	a. Interpret the slope and y-intercept in context of the posed question for the placed line on the scatterplot[10] b. Informally describe correlation by examining how closely the points follow a line c. Predict values of y using values of the independent variable within the range of the sample values of the data set (interpolation) d. Discuss limitation of predictions using values of the independent variable outside of the sample data (extrapolation) e. Discuss the feasability of examining scatterplots in the case of a large data set.	Bivariate data can be displayed in a scatterplot. A relationship is said to be "linear" if the points cluster in an elliptical cloud. When the relationship is linear, the center can be modeled by a line, which can be used for prediction. Error in prediction can be estimated from the size of the residuals, which are measured vertically. Variability from the line is measured by correlation.
Loop 2	How can we extend what we know about the relationship between the mean and the variance to find a criterion for a line of best fit?	a. Determine a method of obtaining the measurements/counts for each variable from each member of the population	a. Find the sample linear regression equation using technology b. Show the residuals on the scatterplot as the vertical distances between the predicted value on the regression line and the actual data point (error in the prediction) with a sign, noting that points above the line have positive residuals and points below the line have negative residuals c. Obtain the residuals through technology and have technology show the residual plot d. Measure the variability in the residuals using sum(y - y-hat)^2 e. Measure the total variability in the values of y	a. Show that the point (xbar, ybar) is always on the regression line b. Write a sample y value as the fitted value + residual c. Show the residuals sum to 0 d. Interpret the residual plot and relate it back to the residuals seen on the scatterplot e. Discuss how the size of the residuals relates to the strength of the association between the two variables. f. Understand how least-squares can be viewed as minimizing the sum of the residuals	Each point can be described as *prediction* plus *residual*. Throughout statistics, variability is measured by the sum of squared distances from the center. The regression line is the line that minimizes the sum of squared residuals.
Loop 3	How can we use the concept of sum of squared residuals to find an interpretation and formula for the correlation? How can bivariate data be analyzed if they aren't linear?	a. Determine a method of obtaining the measurements/counts for each variable from each member of the population	a. Transform the data if necessary b. Find the sample linear regression equation using technology b. Compute the correlation using technology .	a. Understand what correlation represents on the scatterplot b. Understand that r2 is the fraction of the sample variance explained by x c. Relate the correlation to the residuals d. Understand whether causation is determined by the strength of the association e. Understand that transformations can linearize data	Correlation also is a function of the sum of squared residuals ($r^2 = 1 - SSE/SST$) and r^2 can be interpreted as the proportion of variability in the values of *y* that can be accounted for by the relationship of *y* with *x*. Log transformations can linearize data. They are important in many practical applications.

Loop 4[8]	Can we extend the idea of inference from a sample to a population to bivariate data?	a. Determine a data sampling method and obtain a sample from the population of interest b. Discuss whether the data will allow you to estimate causal effects c. Discuss the experimental design of your study	a. Plot the least-squares regression equation from the sample b. Discuss the difference between the population regression model and the estimated sample regression equation, including notation	a. Understand the sample regression equation and correlation as estimates of the population parameters.	The model of a linear relationship between *x* and *y* has the form *response = prediction from true regression equation + random error*. The regression coefficients and correlation from a random sample are estimates of the population parameters.
Loop 5	How much do the estimates of the population parameters vary from sample to sample?	a. Take repeated random samples from the population and compute regression equations	a. Make a plot of the values of the slope from the different samples and compare to the population slope	a. Understand that the slope of the sample regression line has a sampling distribution	An approximate sampling distribution can be constructed by repeatedly drawing random samples from the population to make a distribution of the resulting slopes (or intercepts, or correlations). The estimate of the variability in the slope also is a function of the sum of the squared residuals.

[1]The prerequisite knowledge needed to work through the LP consists of: (1) being able to write the equation of a line and compute the slope and y-intercept, and (2) developing intuition about bivariate data using stacking and color gradients (see Konold & Higgings, 2003 for examples).

[2] At the post-secondary level, regression topics would expand to: multiple regression, other types of linear regression, BLUE, derivation of the

[3] While teaching and learning regression, there are theoretical ideas, historical ideas, definitions, and vocabulary that should be introduced.

[4] The Project-SET LP directly aligns with the GAISE Framework. To illustrate the alignment, the LP is organized around the four GAISE

[5] The Project-SET LP directly addresses the Common Core Standard "Interpreting Categorical and Quantitative Data" S-ID. Notes are made

[6] The Project-SET LP aligns with CCSS S-D.2 " Represent data on two quantitative variables on a scatter plot, and describe how the variables

[7] The Project-SET LP aligns with CCSS S-D.2 " Represent data on two quantitative variables on a scatter plot, and describe how the variables

[8] The Project-SET LP aligns with CCSS S-D.3 "Interpret linear models." In particular, the LP unpacks the bullet "Compute (using technology)

[9] The Project-SET LP aligns with CCSS S-D.3 "Interpret linear models." In particular, the LP unpacks the bullet "Distinguish between

[10] The Project-SET LP aligns with CCSS S-D.3 "Interpret linear models." In particular, the LP unpacks the bullet "Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data."