Straight Talk About Lines

Overview and student objectives

In-Class Activity Length: 25 minutes

Overview

In this in-class activity, students practice identifying explanatory and response variables within two different contexts and are introduced to Least Squares Regression (LSR) analysis to develop their conceptual understanding of the line of best fit.

Throughout the activity, students rely heavily on the *DCMP Data Analysis Tools* to perform all calculations and generate graphs, and they explore a non-linear dataset to demonstrate that not all bivariate datasets should be considered for linear regression analysis.

Objectives

Students will understand:

- The basic concept of the line of best fit.
- The basic concept of the method of linear regression analysis on a given dataset.

Students will be able to:

- Identify the explanatory and response variables given the context of a study.
- Decide when linear regression is appropriate and when it is not appropriate.
- Use data analysis tools to generate appropriate scatterplots and the line of best fit.
- Use data analysis tools to identify the equation of the line of best fit and the correlation coefficient r.

Suggested resources and preparation

Materials and technology

- Computer, projector, document camera
- Student Pages for In-Class Activity
- Practice Assignment
- Excel spreadsheet: *DCMP_STAT_6A_Student_Scores*

Prerequisite assumptions

Students should be able to:

- Identify when a linear regression analysis might be appropriate.
- Identify the explanatory and response variables in a given scenario.
- Calculate the line of best fit and write it using proper notation.

Making connections

This activity:

- Connects back to the distinction between qualitative and quantitative variables, as well as scatterplots and correlation.
- Connects forward to linear modeling, residuals, using linear regression to make predictions, and multiple linear regression.

Suggested instructional plan

Frame the activity (10 minutes)

Resources and Structure	Instructor Suggestions
Think-Pair-Share	Question 1 • In pairs, have students discuss the question and explain to each other what a bivariate dataset is in their own words. Additionally, have students think of examples together to share with the class. • Have a few pairs share their explanatory and response variables. • Take the opportunity to explain that when analyzing bivariate data, it is important to begin by clearly identifying the explanatory and response variables and plotting the data to identify any trends that may be present.
Brief Whole-Class Discussion	• Transition to the in-class activity by briefly discussing the **Objectives for the activity**.

Activity flow (14 minutes)

Resources and Structure	Instructor Suggestions
Technology	• Ensure that every group has access to the *DCMP Data Analysis Tools.*
Small Groups	Questions 2–5 • Combine pairs into groups of four to continue this activity. • As groups are working, circulate to ensure students are identifying variables correctly and assist with technology as needed. • To save time, students can copy and paste the data from the Excel file titled *DCMP_STAT_6A_Student_Scores* to the *DCMP Linear Regression* tool. Alternatively, students can enter the data by hand.

Brief Whole Class Discussion	• Once most groups have finished Question 5, bring the class together for a brief discussion. Have students clearly state and justify their conclusions (answer to Question 5). • Display the line of best fit using a projector and challenge students to visually predict George's final exam score. • Display the five steps of LSR analysis attached to the end of these Instructor Notes. Take a moment to connect the questions students just answered to the five steps involved in LSR analysis. The steps are: o STEP 1: Variable Identification Clearly identify your explanatory and response variables (both quantitative) and then responsibly gather your data. (Make sure samples are random, measurements are accurate, and bias is minimized as much as possible.) o STEP 2: Plot your data Plot your data on a scatterplot, placing the explanatory variable along the x-axis and the response variable along the y-axis. o STEP 3: Check Ensure that the data seem to follow a linear pattern and that there are not some extremely unusual points that perhaps should be excluded from the data. o STEP 4: Line of Best Fit Calculate and draw the equation of this line of best fit ($\hat{y} = a + bx$) and the correlation coefficient (r). Note that the line goes through the data as closely as possible, minimizing how much the data points deviate from the line. o STEP 5: Conclusions Interpret coefficients, assess model accuracy and fit, and make appropriate predictions. • **Anonymity and Confidentiality:** Ask the students if they have any concerns about this dataset. Hopefully students bring concerns about anonymity. For example, Kanye may not want people to know he is a good student. Follow up by asking what the teacher may have done differently (such as use student numbers rather than first names or any tag that can be traced to the actual individual). Questions 6–8 • You may want to rearrange pairs to create new small groups or continue with the same groups. • Continue circulating to ensure students are identifying variables correctly and assist with technology as needed.

	• Monitor Question 6, Part A carefully. This question has the potential to derail the conversations within groups. Be ready to guide students to continue with their analyses regardless of the answers.

Wrap-up/transition (2 minutes)

Resources and Structure	Instructor Suggestions
Wrap-up	• As a wrap-up, have students clearly state and justify their conclusions (answer to Question 8). • Display the scatterplot and discuss the trends. In this case, there is no linear relationship. • Have students refer back to the **Objectives for the activity** and check the ones they recognize. Alternatively, they may check the objectives throughout the activity.
Transition	• In the next activity, we will be learning from a cricket and a scientist who died a long time ago.

Suggested assessment, assignments, and reflections

- Practice Assignment

Steps for Least Squares Regression (LSR) Analysis

Variable Identification: Clearly identify your explanatory and response variables (both quantitative) and then responsibly gather your data. (Make sure samples are random, measurements are accurate, and bias is minimized as much as possible.)

Plot your data: Plot your data on a scatterplot, placing the explanatory variable along the x-axis and the response variable along the y-axis.

Check: Ensure that the data seem to follow a linear pattern and that there are no extremely unusual points that perhaps should be excluded from the data.

Line of Best Fit: Calculate and draw the equation of this line of best fit ($\hat{y} = a + bx$) and the correlation coefficient (r). Note that the line goes through the data as closely as possible, minimizing how much the data points deviate from the line.

Conclusions: Interpret coefficients, assess model accuracy and fit, and make appropriate predictions.

Straight Talk About Lines

1) A teacher wonders if "number of absences per semester" is related to "academic performance" for students in her classes. She looks at her class records from previous semesters and generates a dataset by observing both the final overall average grade and total number of missed classes for each student in a random sample of students. This is an example of a bivariate dataset.

   When working with a bivariate dataset, there are two variables to consider:

   - The explanatory variable (x) is the variable that is thought to explain or predict the response variable of a study. (In previous math classes, this variable may have been referred to as the independent variable.)
   - The response variable (y) measures the outcome of interest in the study. This variable is thought to depend in some way on the explanatory variable. It is often referred to as the "variable of interest" for the researcher. (In previous math classes, this variable may have been referred to as the dependent variable.)

   In this example, the teacher is most interested in how well her students will do in her class, so the response variable is Overall Average Grade. The other variable, Number of Absences, is the explanatory variable.

   Identifying explanatory and response variables can sometimes be difficult. When trying to identify explanatory and response variables, make sure to read the scenario carefully and keep the following phrases in mind:

   Explanatory is used to predict Response.

   (or **calculate)**

   (or **determine)**

   It is good practice to identify both variables and then ask, "Which one is the main outcome or focus of the study?" This variable will be the response variable, and the other variable will be the explanatory variable.

Objectives for the activity

You will understand:

☐ The basic concept of line of best fit.

☐ The basic concept of the method of linear regression analysis on a given dataset.

You will be able to:

☐ Identify the explanatory and response variables given the context of a study.

☐ Decide when linear regression is appropriate and when it is not appropriate.

☐ Use data analysis tools to generate appropriate scatterplots and the line of best fit.

☐ Use data analysis tools to identify the equation of the line of best fit and the correlation coefficient r.

Let's look at an example:

George, a current student, got a 36 out of 50 on the first midterm (C-). He asked his instructor, "If I don't change my study approach, how do you predict I will do on the final exam?"

One way to answer this question is to look at the bivariate data of student scores from a previous class. In this case, we choose a random sample of past students who did not seek out additional tutoring and/or support between the midterm and the final.

The following is a dataset from a random sample of past students who did not seek out advice on study skills or additional tutoring between the midterm and the final exam. To protect their anonymity, only first names are shown.

Student First Name	**Midterm Score (out of 50 points)**	**Final Exam Score (out of 100 points)**
Joe	42	64
Barak	52	94
Hillary	44	87
Donald	25	46
Cher	41	73
Katy	39	73
Taylor	33	53
Miley	40	77
Justin	35	60
Snoop	31	62
Bruno	37	71
Kanye	49	95
Leonardo	38	70
Rosie	45	80
Maya	49	80
Tyra	48	82
Selena	50	81

2) Identify the explanatory and response variables.

Go to the *Linear Regression* tool at https://dcmpdatatools.utdanacenter.org/linear_regression/ and plot the data using the following inputs:

- Under "Enter Data," select "Enter Own."
- Name the *X* (explanatory) and *Y* (response) variables appropriately.
- Copy and paste the data from *DCMP_STAT_6A_Student_Scores* or enter the data in the table by hand. Make sure the explanatory variable is in the first column and the response variable is in the second column.
- Under "Plot Options," select "Regression Line."
- Click "Submit Data" button.

3) Do you think the line of best fit is a good model of the relationship between midterm and final exam score? Explain.

4) Write the equation of the least squares regression line using appropriate notation.

   Part A: Is the relationship positive or negative?

   Part B: What is the value of r? Does this value indicate that the linear relationship between the two variables will be strong, moderate, or weak?

5) Do you think George should be nervous about the final exam?

6) Now, consider the following question: "Can steady driving speed be used to predict fuel efficiency?"

   Part A: If you answered "yes," do you think the relationship between driving speed and fuel efficiency would be positive or negative? If you answered "no," explain.

   Part B: Identify the explanatory and response variables.

7) Go to the *Linear Regression* tool and plot the data using the following inputs:
   - Under “Enter Data,” select “From Textbook.”
   - Under “Choose Dataset,” select “Fuel Efficiency and Speed.”

   Part A: Is the relationship positive or negative?

   Part B: Find the correlation coefficient. Does this value indicate that the linear relationship between the two variables will be strong, moderate, or weak?

8) Is a least squares regression line a reasonable model for the relationship between driving speed and fuel efficiency?

Practice Assignment

1) Which of the following questions could be explored using simple linear regression analysis? Choose all that apply.
   a) Do male pilots tend to have more female children than male children?
   b) Does your reading speed depend on how many hours of sleep you get per night?
   c) Is there a relationship between the weight of horses and how fast they run?
   d) Is there a relationship between the endurance level of horses and the color of their hair?
   e) Are more children born in Spring and Summer compared to Winter and Fall?

2) Dana's Delicious Hot Coco company believes that if they increase advertising on Facebook and other social media outlets, they will increase sales to younger people. The company hires a marketing consultant to conduct a study to see if they are correct. Identify the explanatory variable. Select the best answer.
   a) Costs incurred due to hiring a marketing consultant
   b) Sales revenue generated from young people
   c) Number of ads on social media

3) In the study described in Question 2, identify the response variable. Select the best answer.
   a) Costs incurred due to hiring a marketing consultant
   b) Sales revenue generated from young people
   c) Number of ads on social media

4) An experiment was conducted to see if stressed dogs are more likely to wag their tails to the left when compared to happier dogs. Sixteen dogs were selected for the study. Each dog was placed in a room with distractions and loud noises. The researcher studied the patterns of the dogs' tail-wagging, calculating the percentage of times the dogs' tails wagged to the left. Then, the same dogs were placed in a comfortable room where their owners gently approached them and offered them treats. Again, the researcher studied the patterns of the dogs' tail-wagging, calculating the percentage of times the dogs' tails wagged to the left. Finally, the researcher compared the rates of left-wagging during both the stressful and happy times. They noticed that the left-wagging percentage was higher during the stressful encounters when compared to the happy encounters.

   From the following list, choose the explanatory and response variables.

   Explanatory variable:

   Response variable:

   a) Number of tigers growling at the dogs
   b) Number of treats given to the dogs by the owners
   c) Percentage of times the dogs wagged their tails to the left
   d) Level of stress of the encounter (stressful or non-stressful)

Have you ever wondered what factors make it easier for some cats to be more successful in catching prey? Scientists Michelle Harris and Karen Steudel studied the morphology of 17 domesticated cats.[1] They defined a measure of a cat's ability to catch small animals to be the cat's "take-off velocity," or TOV score (TOV score = how quickly the cat can jump off of the ground). The scientists devised an experiment to see if body mass could be used to predict a cat's TOV score.

[1] Steudel, K. & Harris, M. (2002). The relationship between maximum jumping performance and hind limb morphology/physiology in domestic cats. *Journal of Experimental Biology, 205(24),* 3877–3889. https://doi.org/10.1242/jeb.205.24.3877

5) The following is a small bivariate dataset (continued on the next page) of the TOV scores, in centimeters (cm) per second, and body mass measurements, in grams (g), for the 17 cats involved in the experiment.

Body mass (g)	TOV (cm per second)
3640	334.5
2670	387.3
5600	410.8
4130	318.6
3020	368.7
2660	358.8
3240	344.6
5140	324.6
3690	301.4
3620	331.8
5310	312.6
5560	316.8
3970	375.6
3770	372.4
5100	314.3
2950	367.5
7930	286.3

Part A: Identify the explanatory and response variables in this study. (Make sure to include the units in your description.)

Explanatory variable:

Response variable:

Part B: Go to the *Linear Regression* tool at https://dcmpdatatools.utdanacenter.org/linear_regression/ and plot the data using the "Enter Own" feature.

Do you think it is reasonable to use a linear model to describe the relationship between TOV score and body mass? Explain.

Part C: Use the *Linear Regression* tool to generate the line of best fit.

What is the equation of the line of best fit and the value of the correlation coefficient? Make sure to use proper notation.

Equation:

Correlation Coefficient:

Dr. Suzie Brothers wonders if sodium levels in cereals can be used to predict sugar levels in boxes of cereals. Her stay-home husband Barry wonders if a similar study could use the alcohol content in beer to predict the number of calories in beer bottles. They both collected data, and these data are available in the *Linear Regression* tool at https://dcmpdatatools.utdanacenter.org/linear_regression/.

Suzie's dataset is called "Cereals: Sodium and Sugar," and Barry's dataset is called "Beer Alcohol and Calories."

6) Compare the scatterplots generated by the two datasets to see which study is worth investigating using a simple linear regression model.

Part A: Which study should be abandoned?

a) Suzie's cereal study - using sodium to predict sugar levels
b) Barry's beer study - using alcohol content to predict number of calories
c) Both should be abandoned because neither demonstrate a strong linear relationship.
d) Neither should be abandoned because they both demonstrate that the variables are linearly related.

Part B: Pick the study that has strongly associated variables. Use the information presented in the graph to identify the equation for the line of best fit and the correlation coefficient. Make sure to use proper notation.

Equation:

Correlation coefficient:

Part C: For the equation in Part B, what does represent?

a) The actual number of calories observed
b) The predicted number of calories
c) The actual grams of sugar observed
d) The predicted grams of sugar
e) The actual alcohol content as a percent
f) The predicted alcohol content as a percent
g) All of the above are possible answers
h) None of the above

7) Of the following eight scatterplots, which four are good candidates for linear regression analysis?