INTRODUCTION

In the previous lesson, you predicted the value of the response variable knowing the value of the *explanatory variable* (also known as *predictor variable*) using a best-fit line. So, how do you identify the line that is the best fit? You will use technology to find the equation of the line, but what does it mean to say that a particular line is the best fit? In this lesson, you will investigate this question with the goal of developing a method for determining which line is the best-fit line.

Activity – Part I

Stats textbooks

	price
1	150.67
2	122.00
3	149.10
4	166.15
5	107.95
6	181.95
7	158.95
8	151.95
9	122.00
10	150.67
11	190.95
12	118.95

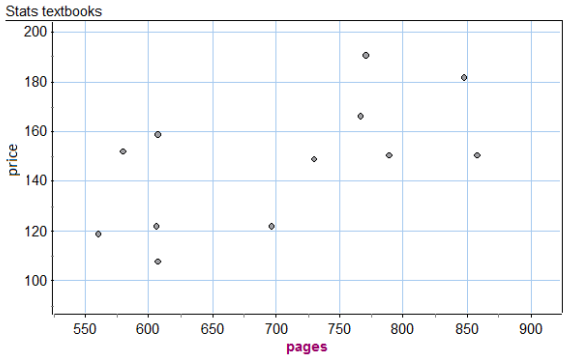
Here are the publishers' suggested list prices in 2010 for 12 popular introductory statistics textbooks. The table below gives the descriptive statistics for the price data.

	Min	Q1	Median	Q3	Max	Mean	Standard Deviation
List price	170.95	122.00	150.67	162.55	190.95	147.61	25.72

1. If someone asks you how much an introductory statistics textbook costs, what prediction would you give? Explain your reasoning.

2. What variables might be useful for predicting the cost of an introductory statistics textbook?

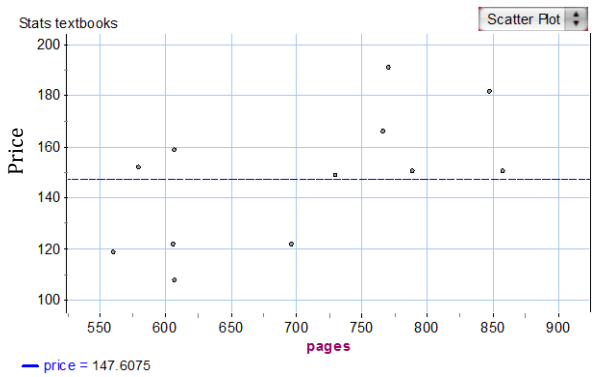
3. The number of pages in the textbook is one variable you could use to predict price. The scatterplot shows the relationship between pages and price for these 12 textbooks. The data have a somewhat linear form and the correlation coefficient is 0.79, so it makes sense to use a line to summarize the relationship between pages and price. Draw a line that you think is a good summary of the relationship between these two variables. Use the graph of your line to predict the price of a 650-page textbook. Then compare your prediction with a classmate.



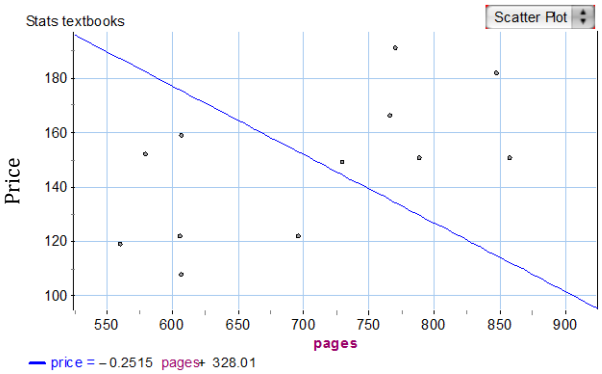
Activity – Part II

Since there are infinitely many lines that you could draw, you need a way to determine which line is the best summary of the relationship between two quantitative variables. You will begin your investigation of how to define a best-fit line by comparing how well four lines predict the list price of the textbooks based on the number of pages.

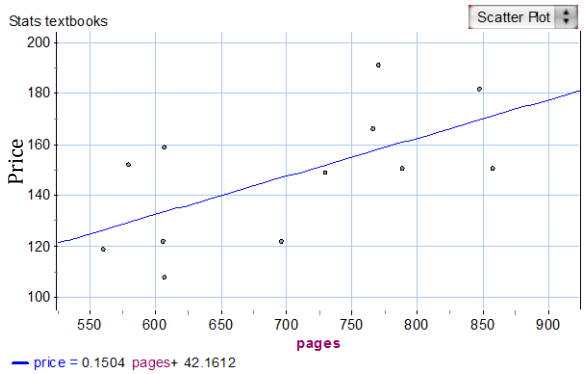
Line A (Mean Price)



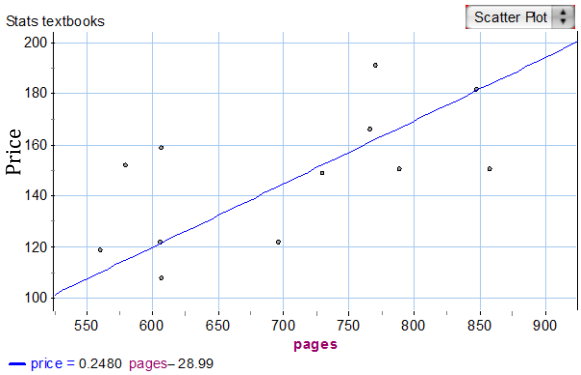
Line B



Line C



Line D



1. Begin by using the equation for each line to complete the two incomplete rows in the table of predicted values on the next page. (You are predicting prices. It makes sense to write prices with two decimal places, such as $147.61 instead of $147.6075 like you see in the table. You might be wondering why you are recording answers to four decimal places. This is because you will need this level of accuracy to develop some ideas later. So, record your answers to four decimal places for these activities.)

Stats textbooks

	pages	price	Line_A_predictions	Line_B_predictions	Line_C_predictions	Line_D_predictions
1	560	118.95	147.6075	187.1700	126.3852	109.8900
2	579	151.95	147.6075	182.3915	129.2428	114.6020
3	606	122.00	147.6075	175.6010	133.3036	121.2980
4	607	107.95	147.6075	175.3495	133.4540	121.5460
5	607	158.95	147.6075	175.3495	133.4540	121.5460
6	696	122.00				
7	730	149.10	147.6075	144.4150	151.9532	152.0500
8	766	166.15	147.6075	135.3610	157.3676	160.9780
9	770	190.95	147.6075	134.3550	157.9692	161.9700
10	788	150.67	147.6075	129.8280	160.6764	166.4340
11	847	181.95				
12	857	150.67	147.6075	112.4745	171.0540	183.5460

2. Which of the four lines do you think results in the best overall predictions of price? Why? How are you selecting the best line?

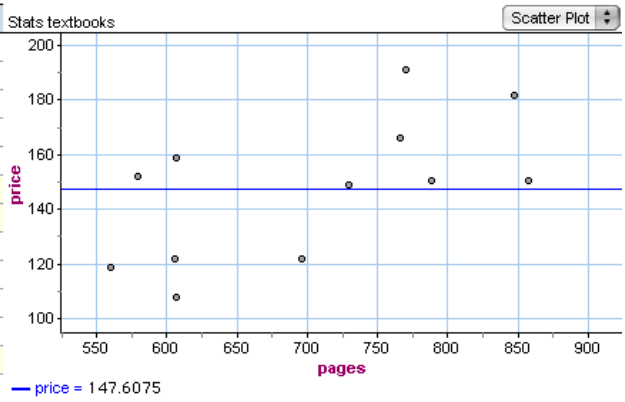
Activity – Part III (Thinking About Prediction Error)

1. For each linear model, complete the missing parts of the table and answer the questions.

Line A (Mean Price)

Stats textbooks_2

	pages	price	Line_A_predictions	Line_A_error
1	560	118.95	147.6075	-28.6575
2	579	151.95	147.6075	4.3425
3	606	122.00	147.6075	-25.6075
4	607	107.95	147.6075	-39.6575
5	607	158.95	147.6075	11.3425
6	696	122.00	147.6075	
7	730	149.10	147.6075	1.4925
8	766	166.15	147.6075	18.5425
9	770	190.95	147.6075	43.3425
10	788	150.67	147.6075	3.0625
11	847	181.95	147.6075	
12	857	150.67	147.6075	3.0625



Identify the following textbooks in the scatterplot and the table:

a). The textbook for which the line comes closest to predicting the list price

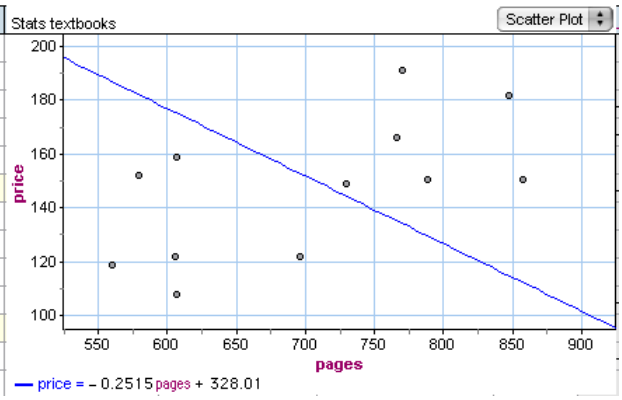
b). The textbook for which the prediction is furthest from the list price

c). In the scatterplot, circle the textbooks that have a negative prediction error. What does a negative error tell you?

Line B

Stats textbooks

	pages	price	Line_B_predictions	Line_B_error
1	560	118.95	187.1700	-68.2200
2	579	151.95	182.3915	-30.4415
3	606	122.00	175.6010	-53.6010
4	607	107.95	175.3495	-67.3995
5	607	158.95	175.3495	-16.3995
6	696	122.00	152.9660	
7	730	149.10	144.4150	4.6850
8	766	166.15	135.3610	30.7890
9	770	190.95	134.3550	56.5950
10	788	150.67	129.8280	20.8420
11	847	181.95	114.9895	
12	857	150.67	112.4745	38.1955



Identify the following textbooks in the scatterplot and the table:

a). The textbook for which the line comes closest to predicting the list price

b). The textbook for which the prediction is furthest from the list price

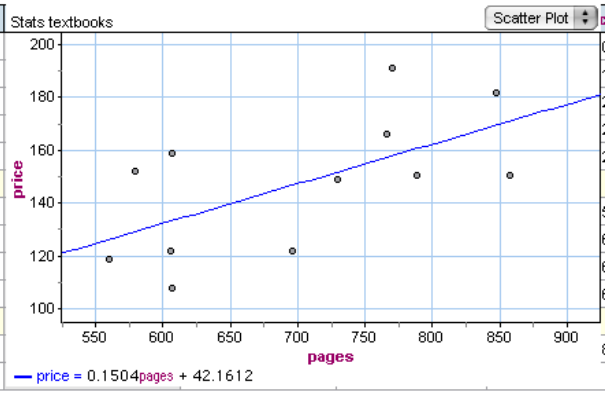
c). How can you tell by looking at the scatterplot if the prediction error for a textbook is positive or negative?

d). Identify a textbook for which Line A predicts too low a price but Line B predicts too high a price.

Line C

Stats textbooks

	pages	price	Line_C_predictions	Line_C_error
1	560	118.95	126.3852	-7.4352
2	579	151.95	129.2428	22.7072
3	606	122.00	133.3036	-11.3036
4	607	107.95	133.4540	-25.5040
5	607	158.95	133.4540	25.4960
6	696	122.00	146.8396	
7	730	149.10	151.9532	-2.8532
8	766	166.15	157.3676	8.7824
9	770	190.95	157.9692	32.9808
10	788	150.67	160.6764	-10.0064
11	847	181.95	169.5500	
12	857	150.67	171.0540	-20.3840



Identify the following textbooks in the scatterplot and the table:

a). The textbook for which the line comes closest to predicting the list price

b). The textbook for which the prediction is furthest from the list price

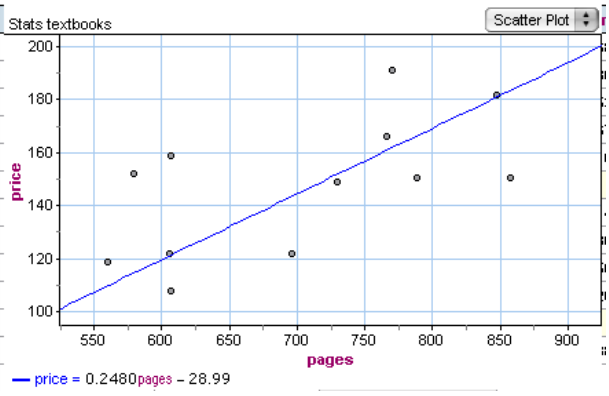
c). All the textbooks for which the predicted list price is within $15 of the actual list price

d). How can you tell by looking at the scatterplot that the prediction error is positive?

Line D

Stats textbooks

	pages	price	Line_D_predictions	Line_D_error
1	560	118.95	109.8900	9.0600
2	579	151.95	114.6020	37.3480
3	606	122.00	121.2980	0.7020
4	607	107.95	121.5460	-13.5960
5	607	158.95	121.5460	37.4040
6	696	122.00	143.6180	
7	730	149.10	152.0500	-2.9500
8	766	166.15	160.9780	5.1720
9	770	190.95	161.9700	28.9800
10	788	150.67	166.4340	-15.7640
11	847	181.95	181.0660	
12	857	150.67	183.5460	-32.8760



Identify the following textbooks in the scatterplot and the table:

a). The textbook for which the line comes closest to predicting the list price

b). the textbook for which the prediction is furthest from the list price

c). all the textbooks for which the predicted list price is exceeds the actual list price by $20 or more

2. The goal is to identify a line that is the best summary of the relationship between pages and price. The best-fit line gives the best predictions of list price, which means that overall it has the least amount of error in the predictions. Rank the four lines from best to worst with the best being the line that gives the best overall predictions of list price. Briefly explain the reasoning behind your rankings.

	pages	price	Line_A_predictions	Line_A_error	Absolute_value_of_Line_A_error	Line_A_error_squared
1	560	118.95	147.6075	-28.6575	28.6575	
2	579	151.95	147.6075	4.3425	4.3425	
3	606	122.00	147.6075	-25.6075	25.6075	655.7441
4	607	107.95	147.6075	-39.6575		1572.7173
5	607	158.95	147.6075	11.3425		128.6523
6	696	122.00	147.6075	-25.6075	25.6075	655.7441
7	730	149.10	147.6075	1.4925	1.4925	2.2276
8	766	166.15	147.6075	18.5425	18.5425	343.8243
9	770	190.95	147.6075	43.3425	43.3425	1878.5723
10	788	150.67	147.6075	3.0625	3.0625	9.3789
11	847	181.95	147.6075	34.3425	34.3425	1179.4073
12	857	150.67	147.6075			

	pages	price	Line_C_predictions	Line_C_error	Absolute_Values_of_Line_C_error	Line_C_error_squared
1	560	118.95	126.3852	-7.4352	7.4352	55.2822
2	579	151.95	129.2428	22.7072	22.7072	
3	606	122.00	133.3036	-11.3036	11.3036	
4	607	107.95	133.4540	-25.5040	25.5040	650.4540
5	607	158.95	133.4540	25.4960	25.4960	650.0460
6	696	122.00	146.8396	-24.8396		617.0057
7	730	149.10	151.9532	-2.8532	2.8532	8.1408
8	766	166.15	157.3676	8.7824		77.1305
9	770	190.95	157.9692	32.9808	32.9808	1087.7332
10	788	150.67	160.6764	-10.0064	10.0064	100.1280
11	847	181.95	169.5500	12.4000	12.4000	153.7600
12	857	150.67	171.0540			

Which measures of the total error help you determine how well a line fits the data?			
Line	*Sum of Errors*	*Sum of Absolute Value of Errors (SAE)*	*Sum of Squares of Errors (SSE)*
A	0.0000	239.0600	7,275.7566
B	–48.9605	485.0950	24,774.1494
C	0.0404	204.6924	4,458.5762
D	32.7460	206.3540	5,734.1069

Statisticians square the errors and then find the line that minimizes the sum of the squared errors. The line that has the smallest sum of the squared errors is called the *least squares regression line*. This line minimizes the sum of the squares of the errors, when compared to **all** other possible lines. You will use technology to find the equation of the least squares regression line.

COMPUTER LAB WORK

	units	textbooks
1	3	120.25
2	4	65.95
3	9	465.00
4	12	430.00
5	14	396.50
6	16	475.00
7	8	208.00
8	1	5.00
9	6	49.10
10	15	685.00
11	9	220.00
12	4	172.00
13	12	302.00
14	12	460.12
15	12	530.00

1 Here you have data collected from students at CCSF in 2009 (table on the right). The variable *units* gives the number of college course units the student reported he or she was taking that semester. The variable *textbooks* gives the amount that the student reported spending on textbooks or other resources required for their courses that semester.

a). Use technology to find the least squares regression line. (Think carefully about which variable is the explanatory variable.)

b). Use the least squares regression line to predict the amount spent on textbooks for a student taking 12 units.

c). Use the regression to predict the amount spent on textbooks for a student taking 10 units.

d). Explain why the least squares regression line is considered the line of best fit.

2 With the following applet, you can draw a line that you think fits the data well and compare your line to the least squares regression line.

www.rossmanchance.com/applets/Reg/index.html

Note: In the applet, errors are called *residuals*. This term comes from thinking about a data point as composed of two parts: the part explained by the regression line (the prediction) and the part that is leftover (called the *residual* or *error*).