Math 1

Linear Regression

Linear Regression

The purpose of linear regression is to capture the relationship between two variables that appear to have a **linear** relationship.

Examples:

The first step in performing regression procedures is to create a **scatter plot** of the data and determine if there appears to be a linear relationship.

This is performed in the calculator by first entering the data in L1 and L2. To do this, press the **[STAT]** key and choose **1:Edit**. Then enter the x-values of your data set into **L1** and the y-values of your data set into **L2**.

Now we are ready to create a scatter plot of the data.

First check to make sure there are no functions entered in **[Y=]**. If there are, clear them.

Now press **[2nd] [Y=]** and choose **1:Plot 1 [ENTER]**. Choose **On** and the **Type:** (scatter plot).

In **Xlist:** type **[2ND] [1]** for L1 and in **Ylist:** type **[2ND] [2]** for L2. (If L1 and L2 are already in Xlist and Ylist, you do not need to type them.

To see your scatter plot, press **[ZOOM] 9:ZoomStat**. Now examine the graph to see if there appears to be a linear relationship. If it is linear, we can find a line of best fit. If the data does not appear linear, a different type of model might be more appropriate.

Example. The table below shows the number of students in Arlington High School along with the years since 2000.

Years since 2000	2	3	4	5	6
Number of stiudents	2892	3042	3087	3136	3296

Enter the data into your calculator and create a scatter plot. Sketch the plot below and be sure to label the axes. Does there appear to be a linear relationship?

Types of Association (Correlation)

The next step in linear regression is determining a linear equation that best fits the data. This line is often referred to as the **"line of best fit"** or the **"least squares regression line."**

To use the calculator to find the equation, press **[STAT] <CALC> 4:LinReg(ax+b)**. This should produce the LinReg Screen.

Xlist: L1

Ylist: L2

FreqList: (leave blank)

Store RegEQ: (use **[VARS] <Y_VARS> 1:Function 1:Y1** to store your equation in Y1=)

Calculate **[ENTER]**

This should produce the regression results screen that shows y=ax+b along with the slope and y-intercept of the regression line. Additionally, the equation has been stored in Y1. If you graph the scatter plot the line will show up on the graph.

Example (cont).

1. Compute a least squares regression line for the previous data set. Write your equation below and clearly label your variables.

2. Interpret the meaning of the slope of the regression line in real life terms.

3. Interpret the meaning of the y-intercept of the regression line in real life terms.

4. Use your regression line to predict the population of Arlington High School in 2012. Round your answer to the nearest whole number. What does this assume about the population of Arlington High School?

5. Use your model to determine between which two years the population is expected to reach 4500.

Interpolation:
Extrapolation:

Application. A real estate agent is trying to determine the relationship between the distance a 3-bedroom home is from New York City and its average selling price. He records data for the 6 homes shown below.

Miles from NYC	10	35	50	65	75	120
Price	755,000	650,000	580,000	505,000	475,000	285,000

a. Create a scatter plot of the data. Sketch your graph below. Label the axes. Does there appear to be a relationship between home price and distance from New York City?

b. Determine a regression model for the data. Write the equation of your model below. Be sure to clearly label your variables.

c. What are the real life meanings of the y-intercept and the slope of your equation?

d. Use your model to determine the predicted price of a home in Woodstock, NY which is located 95 miles from NYC. Would this prediction be *interpolation* or *extrapolation*? Why?

e. Use your model to determine, to the nearest mile, the distance from NYC a home would be if the selling price was exactly $500,000.

Determining How Good a Linear Model Is

Residuals – Obviously when constructing a linear regression model, the actual data points will not be all on the regression line. The distance a point is from the line can be thought of as an error. In other words, how much the model is off from the actual data. This error is called a **residual**.

Residual =

The line of best fit is chosen to minimize the sum of all the residuals. In order to see how good a model fits the data set, a **residual plot** is constructed.

After conducting a linear regression, the residuals will automatically be stored in the calculator. To view the graph of the residuals, the following steps should be taken.

Select **1:Plot 1** and turn it **OFF**

Select **2:Plot 2** and turn it **ON**. Select the **Scatter Plot** icon. In **Xlist:** enter **L1** and in **Ylist:** enter **RESID**.

Choose **[ZOOM] 9:ZoomStat**

A model that fits a data set well will have a residual plot that is **RANDOMLY SCATTERED** which shows no obvious patterns.

Example. Let's return to the New York real estate example.

Miles from NYC	10	35	50	65	75	120
Price	755,000	650,000	580,000	505,000	475,000	285,000

1. Enter the data into L1 and L2 and perform a regression. Write the regression equation below being sure to label the variables.

2. Create a residual plot and sketch the graph below.

3. What does this residual plot say about the model in general?

4. Compute the residual value for the house 35 miles from New York City.

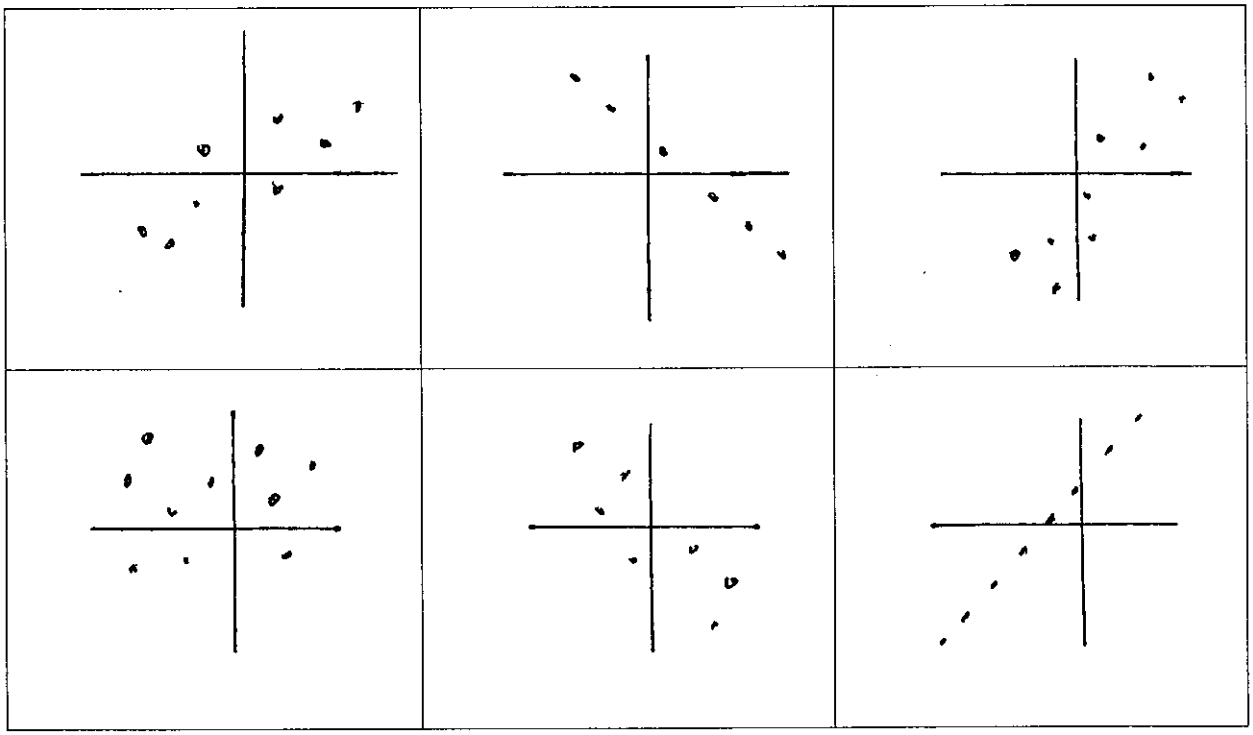
Correlation Coefficient. Another way of determining how good a linear model fits the data is called the **correlation coefficient.** It is denoted by the letter **r**. r measures **how linear a data set is** and has the following characteristics:

1.

2.

3.

4.



To display the r value on the calculator when performing a linear regression, the calculator Diagnostics must be turned on. To do this:

Press **[2nd] [0]** which is CATALOG. Then either press **[D]** and then scroll to **DiagnosticOn** and press **[ENTER] [ENTER]** or scroll from the top to **DiagnosticOn** and press **[ENTER] [ENTER]**.

As long as the Diagnostics are turned on, whenever regression is performed the calculator will provide an r value.

Example. Compute the r value for the New York City example. Interpret the r value in context of the problem.

Applications

1. The following data set represents the yearly movie attendance (in billions) from 2001 to 2008

Year	2001	2002	2003	2004	2005	2006	2007	2008
Attendance	1.44	1.60	1.52	1.48	1.38	1.40	1.40	1.36

1. Enter the data into your calculator and graph it in a scatter plot. Sketch the graph below. Be sure to label the axes.

2. How would you describe the graph? What type of association do you see?

3. Use your calculator to determine the equation of the line of best fit. Write your equation below and clearly label your variables.

4. What is the real life meaning of the slope of the line?

5. What is the real life meaning of the y-intercept of the line?

6. What is the predicted attendance for the year 2010? Show your set up.

7. What year does the equation predict the attendance will surpass 1.8 billion? Round to the nearest year.

8. What is the correlation coefficient for the model? What does this say about the data?

9. Use your calculator to construct a residual plot. Sketch the graph below.

10. What does the plot tell you about how good the model is?

2. The following data set represents the number of stores a company opened from 2000 to 2008

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Stores	241	311	396	519	629	721	809	888	971

11. Enter the data into your calculator and graph it in a scatter plot. Sketch the graph below. Be sure to label the axes.

12. How would you describe the graph? What type of association do you see?

13. Use your calculator to determine the equation of the line of best fit. Write your equation below and clearly label your variables.

14. What is the real life meaning of the slope of the line?

15. What is the real life meaning of the y-intercept of the line?

16. What is the predicted number of stores for the year 2012? Show your set up.

17. What year does the equation predict the 1500 stores will be open? Round to the nearest year.

18. What is the correlation coefficient for the model? What does this say about the data?

19. Use your calculator to construct a residual plot. Sketch the graph below.

20. What does the plot tell you about how good the model is?

1. A fashion designer thought there might be a relationship between the length of a woman's foot and her height. To test this theory, she collected data from 12 women.

Shoe Length (in inches)	Height (in inches)
8.9	61
9.6	61
9.8	66
10.0	64
10.2	64
10.4	65
10.6	65
10.6	67
10.5	66
10.8	67
11.0	67
11.8	70

(a) Sketch a scatterplot of the data.

(b) Write the equation of the line of best fit that models this data. Round to the tenths place.

(c) The designer wants to predict the height of a woman whose shoe length is 7.4 inches. What height does the equation predict?

(d) What does the equation predict for the shoe length of a woman who is 64 inches tall?

(e) Create a residual plot for the data. What does it tell you about the data?

2. The table gives the average hotel rate from 1996 to 2006.

If you plan to stay at a hotel in 2016, what is the rate you would expect to pay?

Equation_____ Rate Expected_____

Year	Rate ($)
1996	70.63
1997	75.31
1998	78.62
1999	81.33
2000	85.89
2001	88.27
2002	83.54
2003	82.52
2004	86.23
2005	90.88
2006	97.78

3. A convenience store manager notices that sales of soft drinks are higher on hotter days, so he assembles the data in the table.

If the high temperature is 91 degrees, what is the predicted number of soft drinks sold?

Equation_____ Expected Cans Sold_____

High Temp	# of Cans Sold
55	340
58	335
64	410
68	460
70	450
75	610
80	735
84	780

4. The accompanying table illustrates the number of passengers (in millions) on ABC Airlines.

Year	Number of Passengers
2000	30.1
2001	33.4
2002	36.2
2003	39.5
2004	42.0
2005	45.8
2006	48.9
2007	52.6
2008	55.3
2009	58.5

(a) Make a scatter plot of the data on your calculator and sketch the graph below. Make sure to label the axes. (Let x=0 represent the year 2000)

(b) Write the linear regression equation for this set of data, rounding values to *five decimal places.*

(c) List the correlation coefficient and explain what it means.

(d) Using this linear regression equation, find the approximate number of passengers, in millions, in the year 2015. Round to the tenths place.

(e) In what year would the number of passengers exceed 70 million?

(f) Create a residual plot and tell what it means about the data.

5. The table shows the number of accidents in the NASCAR Sprint Cup races from 2001 to 2008.

Years	# of Accidents
2001	200
2002	186
2003	235
2004	204
2005	253
2006	237
2007	240
2008	211

(a) Sketch a scatterplot of the data. (Let x=0 represent 2000)

(b) Write the equation for the line of best fit. Round to the tenths place. (remember that x=0 represents 2000)

(c) How many accidents does the equation predict for the year 2005? How does this compare to the actual number of accidents in 2005?

(d) What is the slope of the line? Interpret the meaning of the slope.

(e) What is the y-intercept of the line? Interpret the meaning of the y-intercept.

(f) What is the residual value for the year 2003? How did you figure this out?

Name: _____ **Date:** _____

1. The relationship of a woman's shoe size and length of a woman's foot, in inches, is given in the accompanying table.

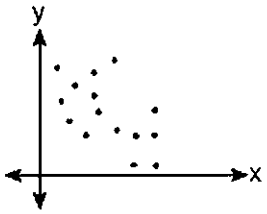
Women's Shoe Size	5	6	7	8
Foot Length (in)	9.00	9.25	9.50	9.75

The linear correlation coefficient for this relationship is

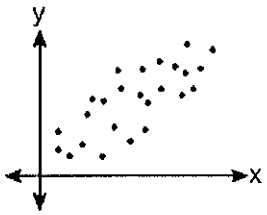
A. 1 B. −1 C. 0.5 D. 0

2. Which graph represents data used in a linear regression that produces a correlation coefficient closest to −1 ?

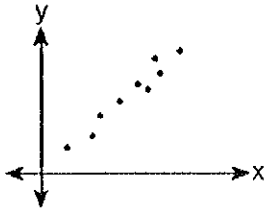
A.



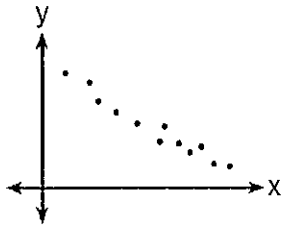
B.



C.



D.



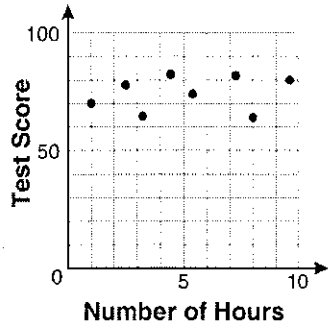
13

3. The accompanying table shows the enrollment of a preschool from 1980 through 2000. Write a linear regression equation to model the data in the table.

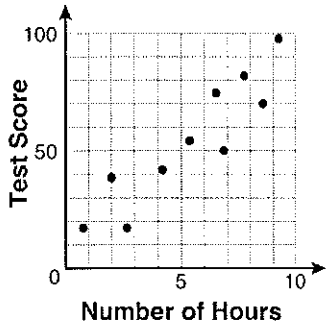
Year (x)	Enrollment (y)
1980	14
1985	20
1990	22
1995	28
2000	37

4. There is a negative correlation between the number of hours a student watches television and his or her social studies test score. Which scatter plot below displays this correlation?

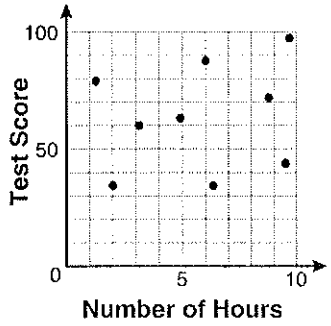
A.



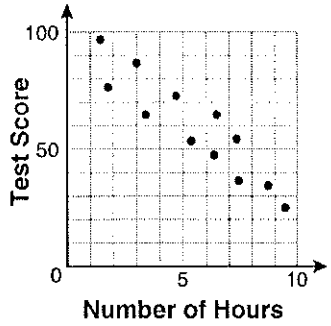
B.



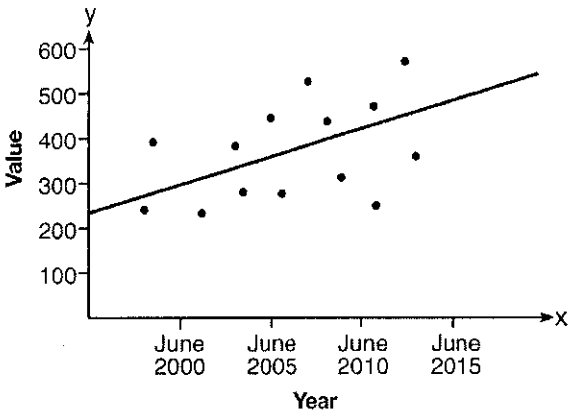
C.



D.



5. Based on the line of best fit drawn below, which value could be expected for the data in June 2015?



A. 230 B. 310 C. 480 D. 540