**Chapter 2.2 – Linear Regression Investigation**

*Part 1: Correlation coefficients*

Correlation is a numeric measure of association between bivariate data. In the table below, sketch a scatterplot given the strength (and direction) of the correlation. (You do not need to provide example explanatory and response variables here, just draw a sketch of the points on the plot.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Strong Positive Correlation** | **Moderate Positive Correlation** | **Weak Positive Correlation** | **No correlation** |
|  |  |  |  |

The correlation coefficient is denoted by the letter *r,* and is always a value between –1 and 1. The following is a set of guidelines for determining the strength of the correlation between two variables:

If $r=1$**,** we say there is a **perfect correlation** between the explanatory & response variables (all points lie

exactly on a straight line)

If $\left|r\right|\geq 0.70$, we say there is a **strong correlation** between the explanatory & response variables

If $0.4 <\left|r\right|<0.70$, we say there is a **moderate** correlation between the explanatory & response variables

If $\left|r\right|<0.40$, we say there is a **weak correlation** between the explanatory & response variables

If $r=0$, we say there is **no correlation** between the explanatory & response variables

\*Note: we use absolute value of *r* because the strength of the correlation doesn’t matter if it’s a positive or negative correlation.

Use this information to guess the correlation of the scatterplots below. Below your guess, write whether the correlation is strong, moderate or weak and the direction (positive or negative).



*r = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ r = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ r = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

STOP!

Check your paper.

*Part 2: Interpreting the linear regression line*

If the scatterplot looks like it follows a linear trend, you can fit a line to the data to model the relationship and make predictions. We will let the calculator do that for us in the next part. For now, we will focus on interpreting linear regression lines.

Gloria sold Girl Scout cookies for several years when she was younger. She kept data on her sales, and determined that the relationship between number of houses visited and number of boxes sold was approximately linear. The scatterplot and regression line are shown below.

Gloria determined the regression equation of this data as follows: $y=1.65x-11.60$

1. What is the explanatory variable (*x)*? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What is the response variable *(y)*? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Interpret the slope: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. According to the regression line, how many boxes will Gloria sell if she doesn’t visit any houses?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Does your answer in part (c) make sense? Why or why not? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. If Gloria visits 60 houses, predict how many boxes of cookies she will sell: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. If Gloria needs to sell 150 boxes to earn a prize, how many houses does she need to visit? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

STOP!

Check your paper.

*Part 3: Calculating the Least-Squares Regression Line (LSRL) using the calculator*

|  |
| --- |
| Table 1: Cellular telecommunications subscribers, 1995-2006 (in millions) |
|  |
| **Year** | **Subscribers (millions)** |
| 1995 | 33.786 |
| 1996 | 44.043 |
| 1997 | 55.312 |
| 1998 | 69.209 |
| 1999 | 86.047 |
| 2000 | 109.478 |
| 2001 | 128.375 |
| 2002 | 140.766 |
| 2003 | 158.722 |
| 2004 | 182.140 |
| 2005 | 207.896 |
| 2006 | 233.041 |
| Source: [*The*](http://www.nass.usda.gov/) *2008 Statistical Abstract* – Table No. 1120 |

 The table below contains data on the total number of cellular telecommunications subscribers from 1995 through 2006 taken from *The 2008 Statistical Abstract*. Using your TI calculator, you will make a scatterplot of the data with a least-squares regression line and determine the strength of the linear relationship. The steps below will guide you through this process.

Graphing a Scatterplot:

**1.** Hit STAT

**2.** Choose “1: Edit…”

**a.** If you have data in your lists, CLEAR them all out.

**b**. Under “L1”, list the years, starting with 1995. Hit ENTER after you input each number.

**c.** Under “L2”, list the number of subscribers (in millions).

 **d**. Hit 2nd MODE (QUIT)

**3**. Hit 2nd Y= (STAT PLOT)

**a**. Hit 1. Make sure the plot is “On.”

**b**. Next to Type: select the first plot (which represents a scatterplot)

**c**. Next to Xlist: if “L1” is not listed, hit 2nd 1 (L1). Next to Ylist: If “L2” is not listed, hit 2nd 2 (L2). This will make it so your graph plots (x,y) values from the lists you made.

**b**. Hit 2nd MODE (QUIT)

**4**. Hit Y=

**a**. Make sure that “Plot1” is highlighted. If it is not, go up and click on it to turn it on.

**b**. Clear out any equations that are stored in any of the rows (scroll all the way down to \Y0=).

**5**. Hit ZOOM 9 (ZoomStat)

Copy the scatterplot you see (be sure to label your axes even though your calculator does not):

Describe the trend you see: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Are the data correlated? If so, in what direction? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Keep going

Now you can use the calculator to find a least-squares linear regression line (LSRL) to fit your data. The calculator will also give you the correlation coefficient for your scatterplot.

1. Hit STAT
	1. Arrow over to “CALC”
	2. Choose “4: LinReg (ax + b)” – *this will calculate your linear function*
	3. Hit ENTER again
2. Your screen should show “LinReg y=ax + b” and then list the values for *a*, *b,*  *r2* and *r.*
	1. What is your value for *a*? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *b*? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
		1. Round these answers to the ten-thousandth’s place
		2. What is your linear regression equation? *y =* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
	2. What is your value for *r2 ?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* *r? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*
		1. Round these answers to the ten-thousandth’s place
		2. r2 is a measure of how well the LSRL fits the data. It’s a percentage. Interpret it by filling in the blank: “The LSRL accounts for \_\_\_\_\_\_\_\_\_\_\_\_% of the variation in number of subscribers.
		3. *r* is the correlation coefficient. Interpret the value by determining the strength & direction of the correlation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. To graph the Linear Regression line, hit Y=
	1. In “Y1=” type in the linear function above.
	2. Be sure “Plot1” is highlighted. If it isn’t, move up and turn it on by hitting ENTER.
4. Hit GRAPH (Your points should cluster roughly around the line)
5. Transfer this graph to the graph below, you may want to use the TRACE button to find exact points to help you graph your line accurately. *Be sure to label both axes and to include both the scatterplot and the linear graph.*



Keep going

Use your LSRL equation to predict the number of cellular telecommunications subscribers in 2013. A well-written response looks like this:

“Assuming the linear trend continues, we predict that in 2013 the number of cellular telecommunications subscribers will be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ million.”

How much faith do you have in your prediction? Explain. A well-written explanation includes a sentence for each of the points below:

* Do you have at least 7 data points? It can be risky to model data with very little data.
* Does the *r2* value indicate a relationship? The closer the *r2* value is to 1, the more variation of the data the LSRL represents and the better the line fits the data. The strength of fit criteria for *r2* follows the correlation strength criteria for *r.*
* Does your regression line fit the shape of your scatterplot? If not, then your predictions might not be very good.
* Does your scatterplot include any outliers, or points well outside the general trend of the rest of the data? If so, these outliers likely have a large effect on the LSRL equation, thus predictions might not be very good.
* Does your prediction make sense? Use your practical, “common sense” knowledge of the problem scenario to determine if you think the prediction makes sense.

Use your LSRL equation to predict the year when the number of cellular telecommunications subscribers was zero. A well-written response looks like this:

“Assuming the linear trend continues into the past, we predict that in the year \_\_\_\_\_\_\_\_\_\_\_\_\_\_ there were no cellular telecommunications subscribers.”

How much faith do you have in your prediction? Explain. *(Hint: focus on the “common sense” criteria.)*

STOP!

Check your paper.

STOP!

Check your paper.