College Algebra - Appendix S

Scattergrams, Regression Lines, and Data Modeling

Mathematical Modeling from Data

Dr. Francesco Strazzullo, Reinhardt University

Objectives

These notes concern topics not covered in our textbook. Screenshots of the TI-84 are generated using the softwares IrfanView4.25 and TI-SmartView2.0 or GeoGebra. We have the following learning objectives.

- Create a scatter plot (or scattergram) of data using a TI-84 or GeoGebra.
- Determine whether a scattergram presents a definite increasing or decreasing behavior.
- Determine whether a scattergram presents a polynomial, a power, an exponential, or a logarithmic behavior.
- Understand what regression lines, regression models, interpolations, and extrapolations are.
- Use technology (TI-84 or softwares) to compute various regression lines and their correlation coefficients.
- Use correlation coefficients to determine the best fit to given data among chosen models (*Data Modeling*).
- Apply data modeling to solve problems and answer questions.

1 Plotting Points: Scattergrams

At this point in the course we know that data relating two measurements can be represented as *ordered pairs*, thus as **points in the Cartesian coordinate system**. Suppose you are given the following situation:

Example 1.1 *Because of the weakening of the U.S. dollar, U.S.-based corporations are generating a growing share of their sales overseas. The following Table 1 shows the percent of sales made abroad in selected fiscal years.*

Year	2004	2005	2006	2008	2009	2010
Percent	32.3	38.9	42.7	43.8	44.7	45.2

Table 1: US-abroad sales for given fiscal years

Therefore Table 1 shows two related measurements: the first one is the fiscal year and the second one is a percent of sales. We can consider x the number of years from fiscal year 2000 and y the percent of sales made abroad, then draw a scatter-plot of these given data. In the softwares using spreadsheets like MSExcel or GeoGebra, we simply copy these data into cells (either by rows or by columns) as seen in Figure 1.

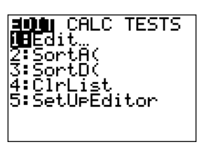
	A	B	C	D	E	F
1	4	5	6	8	9	10
2	32.3	38.9	42.7	43.8	44.7	45.2

Figure 1: Data in a spreadsheet

Note that x-values (reported along the first row) are computed as differences between the actual fiscal year and the "zeroth" year 2000

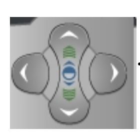
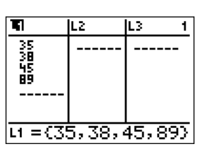
$2004 - 2000 = 4$	$2005 - 2000 = 5$	$\dots$	$2010 - 2000 = 10$

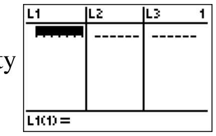
In order to enter in a TI-84 the data from Table 1 in the same form as in a spreadsheet, like Figure 1, we must create "Lists" within the "Stat" window as follows.

1. Type  to enter the "Stat window" 

2. Select "Edit", which by default should be at "1", by typing "1" or moving over and pressing "enter".

3. Now we should see "Lists", labeled L_1, L_2, L_3, and so on. If there is already data stored in a list then we must "clean" or empty it, as follows.

   (a) Move over the list's name by using the directional arrows . Now the list's name should be highlighted 

(b) Type CLEAR then ENTER and the list should be empty 

4. Enter the x-values in L_1 and the y-values in L_2, by typing each and pressing enter, and using directional arrows.

Our data from Example 1.1 will be stored as in Figure 2

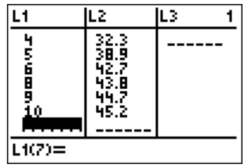
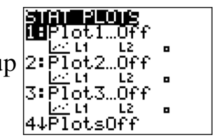


Figure 2: Data in a TI-84

To actually plot the data entered in our lists we must setup the Plot and make sure it is activated. This must be done only once.

1. Type 2nd and Y= to enter the "Stat Plots" setup 

2. Select "Plot1", which by default should be at "1", by typing "1" or moving over and pressing "enter".
3. Move over each option and make the selections as shown here 

4. Here we have the option of "turning-on" the plot, so that it can be graphed, but this can be done in the "y-window, by pressing Y= and selecting "Plot1" so that it is highlighted 

At this point we just have to graph the **Scattergram** for Example 1.1 by typing GRAPH. Most likely we will just see an empty viewing window. That's because we must setup the correct dimensions of our viewing window, which is by default set to $[-10, 10] \times [-10, 10]$. Here we use the standard notation, where the first interval $[-10, 10]$ is the x-range, that is $X_{\text{min}} = -10$ and $X_{\text{max}} = 10$, while the second interval $[-10, 10]$ is the y-range, that is $Y_{\text{min}} = -10$ and $Y_{\text{max}} = 10$. For a "StatPlot" like ours we can use an easy fix. Just type ZOOM and "9", for "ZoomStats". Our scattergram is in Figure 3.

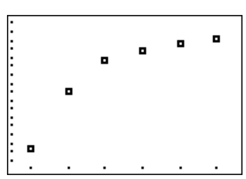


Figure 3: Scattergram in a TI-84

Our example has few measurements. In actual research large data are collected and scattergrams might look "chaotic" like the one in Figure 4.

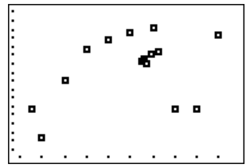


Figure 4: “Chaotic” Scattergram

2 Behavior of a scattergram and graphs of known functions

Our next step is to try and find a function whose graph gets as close as possible to as many as possible points of our scattergram: such graph would be called a **regression line** and this process is called **data modeling**. First, we will try to match graphs of known functions to our scattergram. Looking at Figure 4 one could even argue that those points do not form the graph of a function because they do not pass the vertical line test. Nevertheless we can see that the majority of the points in Figure 4 are distributed from the lower left corner to the upper right corner. We learned in previous chapters that when this happens we have an **increasing function**, or at least the *end-behavior is increasing*, that is for larger x's we have larger y's and for smaller x's we have smaller y's. This is even more visible for the scattergram in Figure 3.

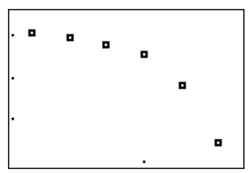
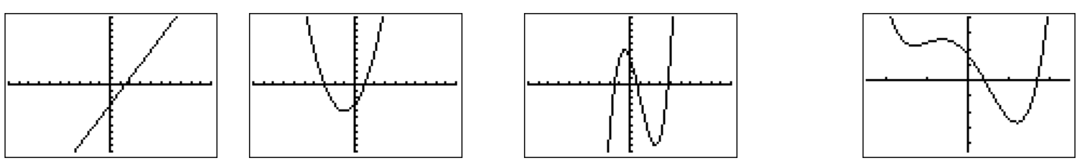


Figure 5: Decreasing Scattergram

The scattergram in Figure 5 could match a **decreasing function**. We want to be more precise and find a function, or more precisely the *equation* of a function, whose graph matches as much as possible a given scattergram. In these notes we will only consider few “known” functions, that we call **models** and report in Figures 6-7.



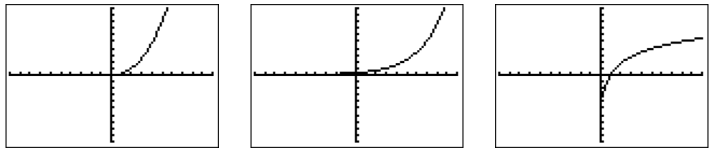
linear	quadratic	cubic	quartic
$y = ax + b,$	$y = ax^2 + bx + c,$	$y = ax^3 + bx^2 + cx + d,$	$y = ax^4 + bx^3 + cx^2 + dx + e$

Figure 6: Polynomial Models

For instance, the scattergram in Figure 3 could match a quadratic (increasing branch of a downward parabola) or a logarithmic model. The best quadratic model matching this scattergram is called *quadratic regression*.

3 Computing Regression Models

The mathematical process for finding **regression models** is beyond the learning objectives of this course: we content ourselves with learning how to compute regression models by using technology, namely a TI-84 (although softwares like MSExcell or Geogebra, for instance, could be used). Different types of regression lines can be computed, according to the models shown in Figures 6-7: LinReg, QuadReg, CubicReg, QuartReg, PWRReg, EXPReg, and LNReg. Some are better fits than others: the **correlation coefficient** r^2 (or R^2) measures of how good a regression line is, that is how close the graph is to the majority of points in a scattergram. The correlation coefficient is usually denoted by



power
$y = ax^b$,

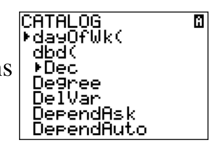
exponential
$y = ab^x$,

natural logarithmic
$y = a + b\ln(x)$

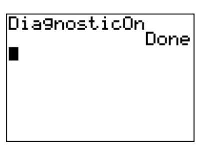
Figure 7: PWR, EXP, and LN Models

r (or R), but a more accurate measurement is given by r^2 and by construction one has $0 < r^2 \leq 1$. The closer r^2 is to 1 the better the regression line is: when $r^2 = 1$ then we say that we have a **perfect fit**, since all the points of the scattergram are part of the regression line (the graph of the regression model). Remember that there is a unique n-th degree polynomial going through $n+1$ given points: for instance there is a 5-th degree polynomial that is a perfect fit to the scattergram in Figure 3.
Why do we need a regression line? We actually use the regression model, that is the equation whose graph is the regression line. By using this equation we can *estimate measurements we are not able to take*, for example because data were not available or are yet to be available. When we estimate outputs for inputs within the given domain, we are computing an **interpolation**. For instance in Example 1.1 we do not know the percent of sales abroad during fiscal year 2007 and we can use a regression model to estimate what that percent could have been. On the other side, we compute an **extrapolation** when we use a regression model to estimate the outputs for inputs outside the given domain. For instance an extrapolation in Example 1.1 would be the percent sale in fiscal years 2011 or 2002. These estimations are accurate only as much as the regression line is and only as close as possible the input is to the given domain. For instance no regression model could be effectively accurate to estimate what the percent of abroad sales was in 1995 or will be in 2025!

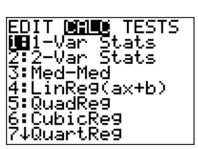
Now we describe how to compute and use regression models. First, a "StatPlot" (scattergram) must be already setup. Second, we must "turn-on" the application that computes correlation coefficients:

1. Type 2nd and 0 to enter the "Catalog" of applications, where we need to "turn-on" the "Diagnostic".
2. Type 2nd and x^{-1} to move to the "D" page of the list of applications 

3. Move down with directional arrows until "DiagnosticOn" is selected 

4. Type ENTER twice to see this screen 

Now we are ready to compute regression models.

1. If you are not already there, go to the calculating window by typing 2nd and MODE.
2. Type STAT and move over "Calc" 

3. Select the required regression line, by moving over it and typing .
4. Now we should be in the calculating window. By pressing "enter" again the regression line will be computed and on the screen we will see all the parameters needed to write its equation, moreover the last parameter should be the correlation coefficient r^2.

Let's see these steps at work.

Example 3.1 *Consider the data from Example 1.1, then let x be the number of years from fiscal year 2000 and y the percent of sales made abroad by U.S.-based corporations. The scatter-plot for this data is shown in Figure* **??**.

1. *Compute the* **quadratic regression** *for this data and report your unswear rounded to the fourth decimal place.*
2. *Use the (rounded) quadratic regression to interpolate the percent of abroad sales made during the fiscal year 2007 and to extrapolate that during fiscal year 2011.*

[**Solution 3.1**] The given data is already stored in L_1 and L_2.

1. In Figure 8 we report the screens that appear after each of the above steps and the sequence of keys typed.

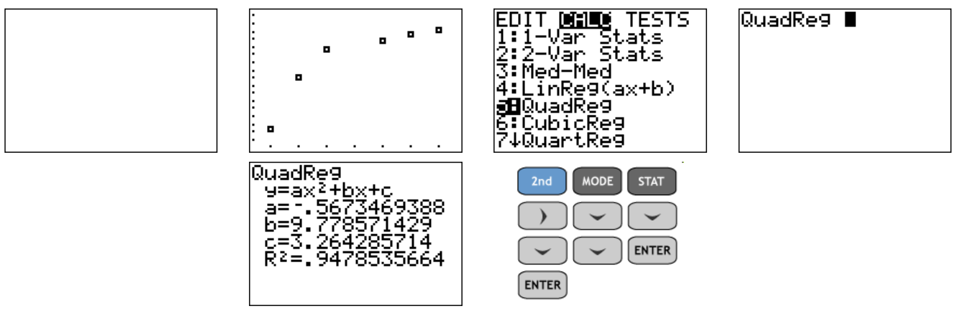


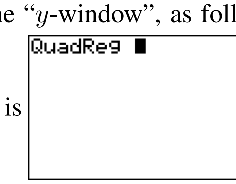
Figure 8: Quadratic Regression for Example 3.1

We need to copy these parameters with the required approximations: in this way we obtain the **rounded model** in equation (1).

$$y = -.5673x^2 + 9.7786x + 3.2643, \qquad r^2 = .9479 \tag{1}$$

2. We must evaluate the function (1) for the given years 2007, which corresponds to $x = 2007 - 2000 = 7$, and 2011, which corresponds to $x = 2011 - 2000 = 11$.
   (a) Interpolation: $y = f(7) = -.5673(7)^2 + 9.7786(7) + 3.2643 = 43.9168$ percent.
   (b) Interpolation: $y = f(11) = -.5673(11)^2 + 9.7786(11) + 3.2643 = 42.1856$ percent.

We also have the option to store the original **unrounded model** in the "y-window", as follows (screenshots are reported in Figure 9).

1. Start from the previous Step 4, before typing , when the screen is 

2. Type  and move over "Y-Vars".
3. Type  to select "Y-Vars".
4. Now we are in the "Function" list, where we can decide where to store an expression in the "y-window": we must move over any of Y_1, Y_2, and so on, then press "enter" (assume we select Y_1).
5. Finally press enter again. The unrounded model is now stored in the "y-window" for Y_1.

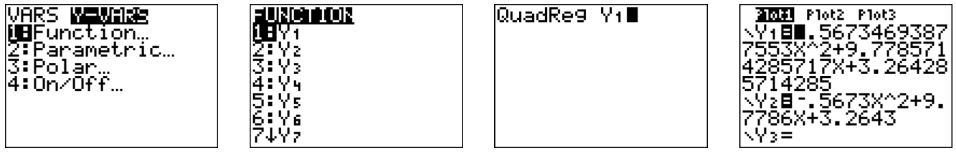


Figure 9: Unrounded model for Example 3.1

Note that the last screenshoot in Figure 9 shows both the unrounded model in Y_1 and the rounded one in Y_2. Evaluating Y_1 at the given years we find

1. Interpolation: in 2007 $y = 43.9143\%$.

2. Interpolation: in 2011 $y = 42.1796\%$.

These values are pretty close to those of the unrounded model and the graphs of the two models show that, being almost undistinguishable (see Figure 10).

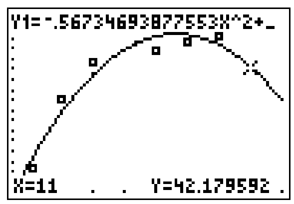


Figure 10: Quadratic Regression lines for Example 3.1

4 Finding the best fit

Now we are ready to compare various regression models and choose the one that best fits our scattergram, that is the one whose correlation coefficient is larger (and closer to 1).

Example 4.1 *Consider Example 3.1.*

1. *Compute the quadratic, the cubic, and the logarithmic regressions, rounded to the fourth decimal place, and state which one is the best fit to the given data.*

2. *Use the best fit to estimate the percent of abroad sales made during the fiscal year 2007.*

[**Solution 4.1**] The quadratic regression and its correlation coefficient are reported in equation (1).

$$y = -.5673x^2 + 9.7786x + 3.2643, \qquad r^2 = .9479$$

1. The solutions computed with the TI-84 are reported in Figure 11. We stored the models in Y_1, Y_2, and Y_3.

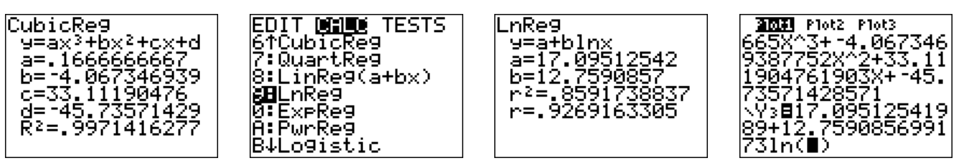


Figure 11: More Regression Models for Example 4.1

The cubic regression model is

$$y = .1667x^3 - 4.0673x^2 + 33.1119x - 45.7357, \qquad r^2 = .9971 \tag{2}$$

The natural logarithmic regression model is

$$y = 17.0951 + 12.7591 \ln(x), \qquad r^2 = .8592 \tag{3}$$

The largest correlation coefficient r^2 is the one in equation (2), therefore the best fit (among these models) is the cubic regression.

2. Interpolation: $y = f(7) = .1667(7)^3 - 4.0673(7)^2 + 33.1119(7) - 45.7357 = 43.928\%$.

5 Recap Example

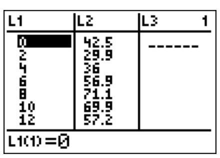
Example 5.1 *Table 2 gives the number of births, in thousands, to females over the age of 35 for a particular state every two years from 1970 to 1986. Use technology to answer to the following questions. Report your answers rounded to the fourth decimal place.*

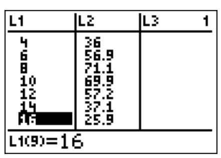
Year	1970	1972	1974	1976	1978	1980	1982	1984	1986
Births (thousands)	42.5	29.9	36.0	56.9	71.1	69.9	57.2	37.1	25.9

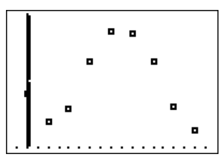
Table 2: Births to female over age 35 for given years

1. *Draw a scattergram for the given data, where x is the number of years after 1970, and establish its end-behavior.*
2. *Find the quartic and the exponential function that are the best fit for these data.*
3. *According to the best model, how many births were there to females over the age of 35 in this state in 1987?*

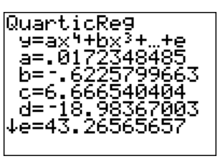
[**Solution 5.1**] Here we just report the screenshoots.

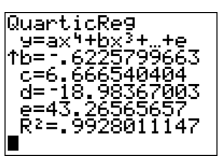
1. 

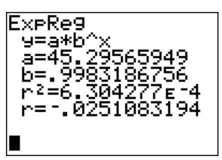




The scattergram seems to fit a quartic model.

2. 





The exponential model is really unfit, with $r^2 = .0006$, therefore the quartic one is the best fit

$$y = .01723x^4 - .6226x^3 + 6.6665x^2 - 18.9837x + 43.2657, \qquad r^2 = .9971$$

3. Extrapolation: for 1987 we use $x = 17$, then $y = f(17) = 27.9098\%$.