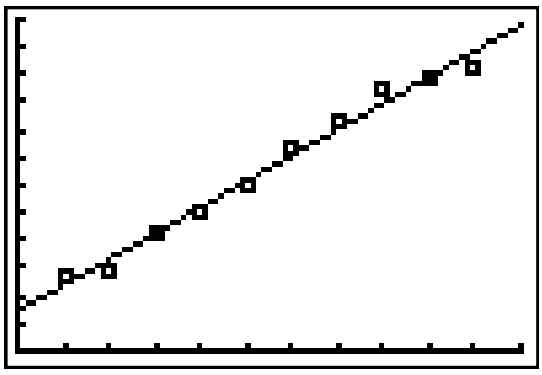
CHAPTER

5



Regression

5.1	Least-Squares Regression
5.2	Residuals and Influential Observations
5.3	Common Errors
5.4	Selected Exercise Solutions

Introduction

In this chapter, we show how to find the least-squares regression line through the data, look at the coefficient of determination, and explore influential observations. Lastly, we show how to work with the residuals of the regression line.

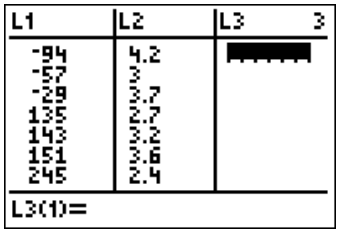
5.1 Least-Squares Regression

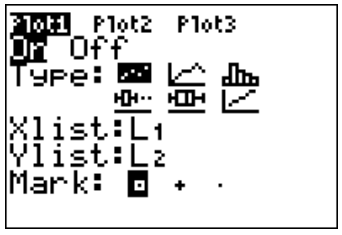
In this section, we will compute the least-squares line of two quantitative variables (the line found by *minimizing* the *squared* errors in the model) and graph it through the scatterplot of the variables. We will also use the line to predict the *y*-value that should occur for a given *x*-value.

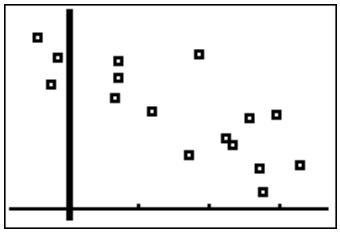
Example 5.1 Does fidgeting keep you slim? The data from Example 5.1 are repeated below. Graph the data and describe the relationship.

NEA (cal)	-94	-57	-29	135	143	151	245	355
Fat (kg)	4.2	3	3.7	2.7	3.2	3.6	2.4	1.3
NEA (cal)	392	473	486	535	571	580	620	690
Fat (kg)	3.8	1.7	1.6	2.2	1	0.4	2.3	1.1

Solution. Graphing the data should always be the first step in fitting any model. If the data are not reasonably linear, fitting a straight line model is a worthless exercise. We have entered the NEA values in L1 and the fat gain in L2. Define the scatterplot (we want to model weight gain based on non-exercise activity) and press [ZOOM] [9].



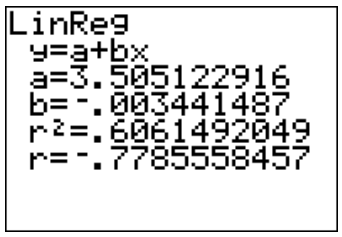


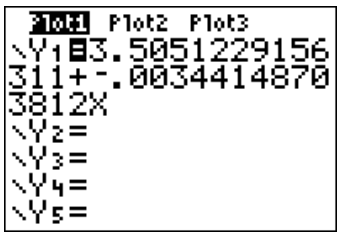


We can see the relationship is negative (decreasing) and fairly linear. The relationship is fairly strong (the linear pattern is clear), but the correlation will not be very close to –1.

Example 5.2 Using a regression line. Fit the regression line to the data from Example 5.1, and show it on your data graph. What weight gain would you predict for a person who increases NEA by 400 calories?

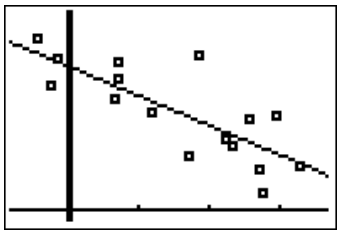
Solution: We used the LinReg(a+bx) command from the [STAT] CALC menu in Chapter 4. We use this again but save the equation in a Y variable so that we can graph it. To add storing the regression function in Y1, press [VARS], [▸] to Y-VARS, [ENTER] for Function and [ENTER] for Y1. If you press [Y=] you'll see the regression equation.



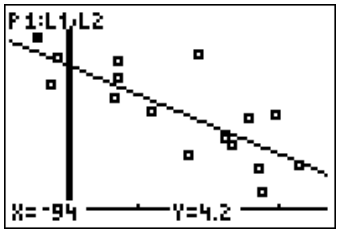


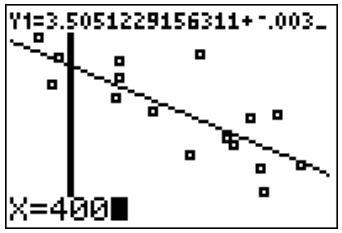
We have the regression function Fatgain (kg) = 3.505 – 0.0034*NEA. The correlation is –0.778. This regression function tells us that for every calorie of NEA activity, weight gain decreases by about 0.0034 kg. While not a lot (the coefficient is close to 0), the relationship is fairly strong.

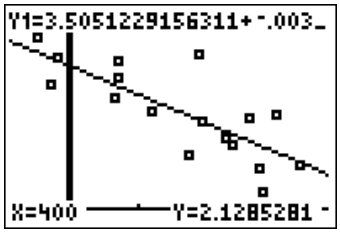
To see the regression line through the data, you simply need to press GRAPH (assuming you just graphed the data as shown above). Note the regression line doesn't seem to go through any of the actual data points, but does go through the middle of the data.



With the equation of the regression line computed and displayed on the graph as above, we can use the TRACE (F3 on a TI-89) to evaluate the function for a specific *x* value. Press TRACE. At the upper right, the active plot is displayed. This indicates we are tracing the scatterplot. Press ▼ to switch to the line. The top line will change to display the equation of the line. Type in 400 (NEA change we want to find the weight gain for) and press ENTER. We see that a NEA change of 400 calories gives a predicted weight gain of 2.129 kg.

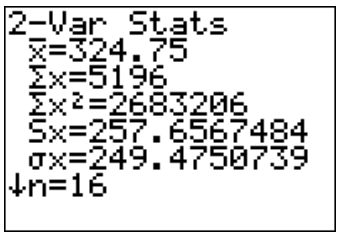


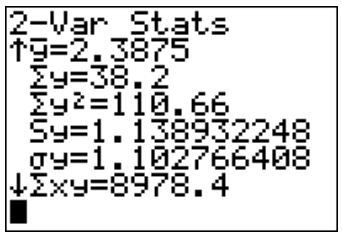


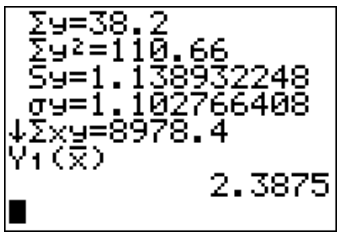


Alternately, we can access the Y1 function from the Home screen. To do so, press VARS, arrow right to Y-VARS, enter 1 for Function, enter 1 for Y1, then enter the command Y1(400). Press ENTER to perform the calculation.

We can also verify that the point $(\bar{x}, \bar{y})$ is on the regression line; but first we must compute the statistics. We can do so simultaneously with the 2-Var Stats command from the STAT CALC menu because the two data sets are the same size. Enter the command 2-Var Stats L1,L2. Then enter Y1($\bar{x}$) by recalling $\bar{x}$ from the VARS Statistics menu.







Example 5.4 Using r^2. With the calculator's diagnostics turned on, the LinReg(a+bx) command also displays the values of r and r^2. In this case, r^2 = 0.606. Thus, 60.6% of the variation in weight gain among these subjects is explained by the straight-line relationship with non-exercise activity changes.

5.2 Residuals and Influential Observations

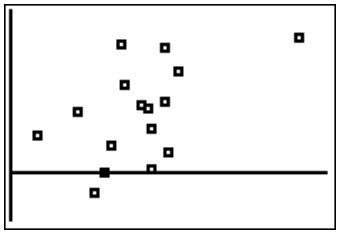
We now complete an exercise to demonstrate how to work with the residuals (errors in the model, defined as $e_i = y_i - \hat{y}$) of a least-squares regression line. These are computed for each regression and stored in a list called RESID. Ideally, these will be randomly distributed around the x-axis ($y = 0$ line). Any indications of pattern (curvature, widening (or narrowing) as x-values increase) indicates a violation of assumptions for the regression. They can also be used to help identify outliers and potentially influential points in a regression.

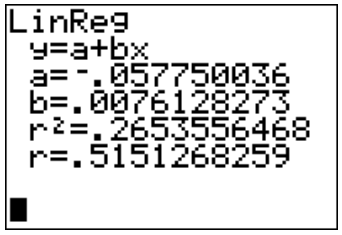
Example 5.5 I feel your pain. Empathy means beng able to understand what others feel. To see how the brain expresses empathy, researchers recruited 16 couples in their midtwenties who were married or had been dating for at least two years. They zapped the man's hand with an electrode while the woman watched, and measured the activity in several parts of the woman's brain that would respond to her own pain. Brain activity was recorded as a fraction of the activity observed when the woman herself was zapped with the electrode. The women also completed a psychological test that measures empathy. Will women who score higher in empathy respond more strongly when their partner has a painful experience? We want fit the model, obtain the residuals, and create a scatterplot of the residuals against our x (predictor) variable.

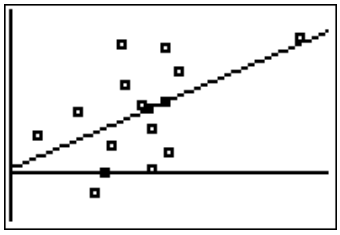
Subject	1	2	3	4	5	6	7	8
Empathy score	38	53	41	55	56	61	62	48
Brain activity	–0.12	0.392	0.005	0.369	0.016	0.415	0.107	0.506
Subject	9	10	11	12	13	14	15	16
Empathy score	43	47	56	65	19	61	32	105
Brain activity	0.153	0.745	0.255	0.574	0.21	0.722	0.358	0.779

Solution. We use the empathy score as the x (predictor) variable in this model. The data have been entered in L1 (empathy) and L2 (brain activity). We first define a scatterplot

of the data and press [ZOOM][9] to display it. The relationship is fairly linear, but there is an outlier at the upper right. We fit the model as shown below.



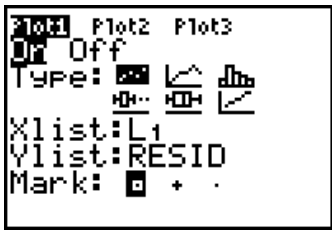


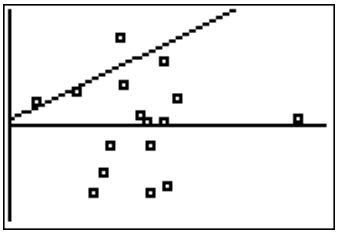


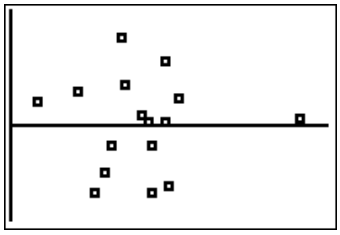
We have a regression equation of BrainActivity = –0.058 + 0.0076*Empathy. The correlation is 0.515. This model (empathy score) explains 26.5% of the observed variation in brain activity.

We now proceed to examine the residuals by plotting them against our predictor variable (empathy score). Define a scatterplot as shown below. To access the list named RESID, press [2nd][STAT], and locate the list name. Press [ENTER] to transfer the list name to the plot definition screen. Press [ZOOM][9] to display the plot. Oops! The regression line shown on our plot below is *not* part of the residuals plot! Go to the [Y=] screen and press [CLEAR] to erase the equation. Redisplay the plot using either [ZOOM][9] or [GRAPH].

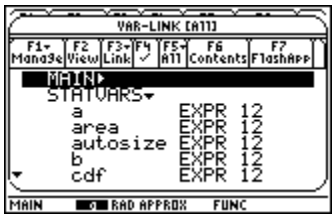
This plot is pretty much ideal, except for that outlier already seen on the original plot of the data. With the exception of that point, the rest of the points are relatively evenly spread around the $y = 0$ axis and seem randomly distributed.

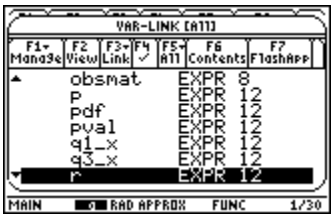


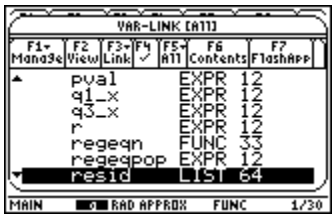




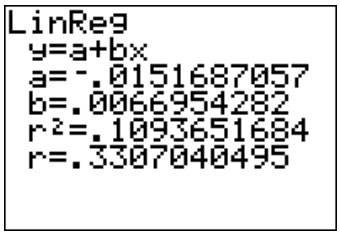
TI-89 Note: To locate the resid list on the [VAR-LINK] screen, press ◀ to collapse the MAIN folder. Press the down arrow to get inside the STATVARS folder then press [2] (R in alpha mode) to get to the portion of these variables that begin with the letter r. Arrow down until you find the list and press [ENTER] to select it and transfer this name to the plot definition.







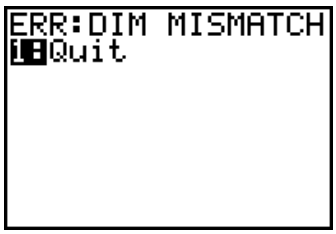
Example 5.6 An influential observation? Subject 16 in the preceding example is an outlier. Is this observation influential? An influential point is one that changes the regression function.



Solution. Deleting subject 16 should lower the correlation (r) and coefficient of determination (r^2), since its presence in the data visually strengthens the relationship. Delete both the values for that subject (otherwise, you'll have a DIM MISMATCH error) and recompute the regression. Notice that r has decreased from 0.515 to 0.331; empathy now explains only 10.9% of the variation seen in brain activity.
The slope didn't change much (from 0.0076 to 0.0066 – a change of about 13%). This subject is not influential to the relationship. Returning to the graph of the data with the regression line, this subject is close to the line.

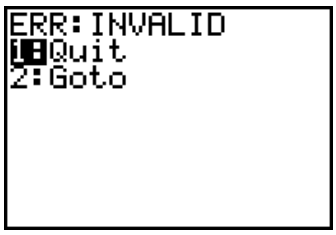
5.3 Common Errors

Err:Dim Mismatch



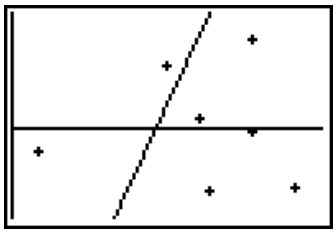
We've seen this one before. Press ENTER to quit. This error means the two lists referenced (either in a plot or a regression command) are not the same length. Go to the STAT editor and fix the problem.

Err: Invalid



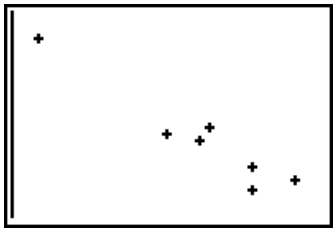
This error is caused by referencing the function for the line when it has not been stored. Recalculate the regression being sure to store the equation into a y function.

What's that weird line?



This error can come either in a data plot (an old line still resides in the Y= screen) or the stored regression line is showing in the residuals plot, as shown here. As stated earlier, the regression line is *not* part of the residuals plot and shows only because the calculator tries to graph everything it knows about. Press Y= followed by CLEAR to erase the unwanted equations, then redraw the graph by pressing GRAPH.

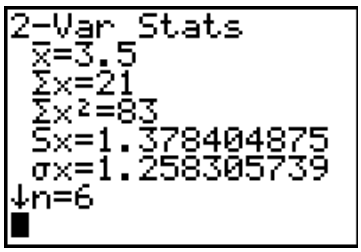
This doesn't look like a residuals plot!

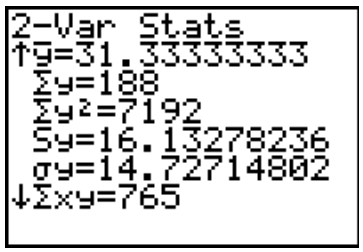


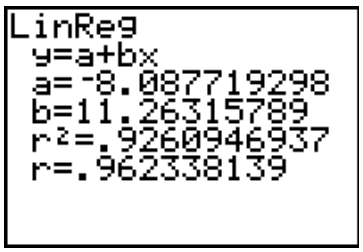
It doesn't. Residuals plots *must* be centered around $y = 0$. This error is usually caused by confusing which list contains the y's and which the x's in computing the regression or in defining the residuals plot. Go back and check which list is which, and recompute the regression, or redefine the plot.

5.4 Selected Exercise Solutions

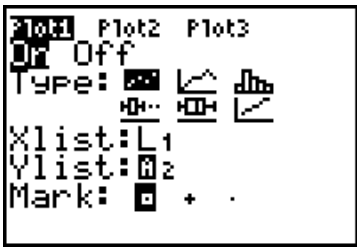
5.3 The data have been entered in L1 (distance) and L2 (days). The relationship is linear. Since there are an equal number of observations in each list, we use [STAT], CALC, 2:2-Var Stats to calculate the summary statistics. To find the correlation (and regression coefficients) use [STAT], CALC, 8:LinReg(a+bx). To "hand calculate" the slope, we have $b = r\frac{s_y}{s_x} = .9623\frac{16.133}{1.378} = 11.266,$ which agrees to within rounding error. Similarly, we find the intercept as $a = \bar{y} - b\bar{x} = 31.333 - 11.266 * 3.5 = -8.098,$ which also agrees (to within rounding errors).

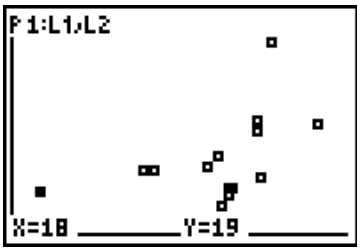




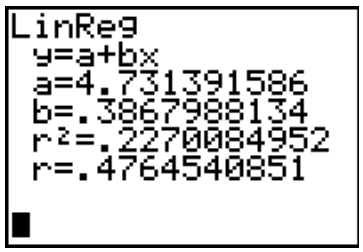


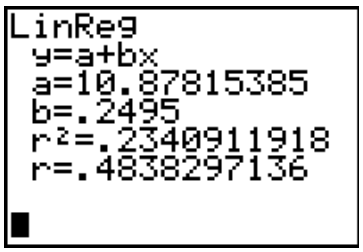
5.11 The data have been entered in L1 (Outsource percent) and L2 (Delay percent). To create a scatterplot of these data, define the plot using the [2nd][Y=] (STAT PLOTS) menu as Plot1, shown below. Press [ZOOM][9] to display the plot.





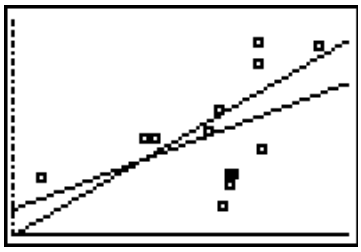
We'll find the regression equations (and correlations) with and without Hawaiian Airlines. Compute both regression lines using [STAT], CALC, 8:LinReg(a+bx). To store the equations, press [VARS], arrow to Y-VARS, press [ENTER] for 1:Function and [ENTER] for 1:Y1 for the first line; use Y2 for the second line, after deleting Hawaiian Airlines (80, 70).

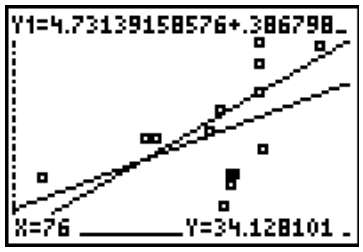


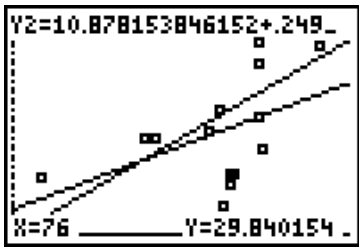


Hawaiian was influential for the regression, but not for the correlation. The correlation increased from 0.476 to 0.483. The regression equation changed from DelayPct = 4.73 + 0.387*OutsourcePct to DelayPct = 10.878 + 0.250*OutsourcePct.

Now, press GRAPH since there is no need to resize the window. Both regression lines will be added. Press TRACE. You will be on the scatterplot, as indicated by P1 in the upper left corner. Press the down arrow to move to Y1. Type in 76 and press ENTER. The equation that included Hawaiian Airlines predicts 34.1% delayed flights. Press the down arrow again and reenter 76. The prediction from the regression that did not include Hawaiian is 29.8% delayed flights – a decrease of 4.3%.







Continue your practice with these exercises:

5.35 Keeping water clean.
5.37 Our brains don't like losses.
5.39 Managing diabetes.
5.41 Managing diabetes, continued.
5.51 Beavers and beetles.
5.55 Saving energy with solar panels.