# Chapter 5
## Summarizing Bivariate Data

### 5.0 Introduction

In Chapter 5 we address some graphic and numerical descriptions of data when two measures are taken from an individual. In the typical situation we are interested in the question of whether two variables are somehow related, and whether or not the nature of that relationship is linear. That is, can we describe the typical behavior of the variables in the manner of a common algebraic straight line, $y = mx + b$? Another description of the data will be numeric – to what extent do our actual data points lie along our straight-line? Our summarizing line is, the least squares best fit line, and our numeric description of the degree of "fit" of the line to our data points is Pearson's correlation coefficient. We also assess visually the "goodness" of our fit of the line to the data by appealing to the residual plot.

And – wonder of wonders! -- the TI-83 will do it all. Let's analyze the data of Example 5.4, the relation between foal weight and mare weight.

### 5.1 Pearson's correlation

**Example 5.4: Is Foal Weight Related to Mare Weight?**

Foal weight at birth is an indicator of health, so it is of interest to breeders of thoroughbred horses. Is foal weight related to the weight of the mare (mother)?

| Observation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Mare weight (*x*, in kg) | 556 | 638 | 588 | 550 | 580 | 642 | 568 | 642 | 556 |
| Foal weight ( *y*, in kg) | 129 | 119 | 132 | 123.5 | 112 | 113.5 | 95 | 104 | 104 |

| Observation | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|
| Mare weight (*x*, in kg) | 616 | 549 | 504 | 515 | 551 | 594 |
| Foal weight ( *y*, in kg) | 93.5 | 108.5 | 95 | 117.5 | 128 | 127.5 |

We begin, as always, by entering the data after the **Stat** > **`Edit`** sequence. Remember, this data is bivariate so you will have to enter the data in two separate lists. Just to keep you on your thoroughbred toes, we will use List5 and List6. After entering your data in whatever lists you choose, execute the sequence,

**Stat** > **`Edit`** > **`Calc`**…

Take a deep breath, and check out these options:

```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9:LnReg
0:ExpReg
A:PwrReg
B:Logistic
C:SinReg
```

If there were any justice in the world, one would pick `2:2-Var Stats` from the list, and out would pop the Pearson correlation coefficient. Unfortunately, for reasons known only to the TI-83 design engineers choosing that option gives you all the information you need if you wanted to calculate Pearson's correlation using the formula in Exercise 5.13 – but doesn't give you the correlation!

To get Pearson's *r* you have to choose a different option, one that is not the most obvious choice. (If you have read Section 5.2 in POD you will know why this is a reasonable choice, but it still isn't obvious.) The lack of obvious is more than compensated for by the fact that you have two options that are equally adept at presenting the correlation coefficient: `4:LinReg(ax+b)` and `8:LinReg(a+bx)`. Both these options accomplish the same thing, but they use the variables *a* and *b* in different roles. In POD the variables are used thus: $y = a + bx$. It is probably better to use the choice that matches POD but either way the calculator will give the same numeric values.

Now we have some bad news to give you: (a) picking either of these choices will get you information you haven't asked for, and (b) you may not actually get the correlation you are hoping for. But don't lose hope yet – we'll surmount every impediment to success and deliver *r* for your consideration. Choose the following sequence:

**Stat** > **Calc** > **`8:LinReg(a+bx)`** **>** $L_5$, $L_6$ **>** **`Enter`**

and let's see how lucky we are. (If your data is anywhere except List1 and List2, you have to tell the TI where they are – hence you need to explicitly add the $L_5$, $L_6$ .) Here's what we see on our calculator:

```
8:LinReg(a+bx)  L5, L6
```

and then…

```
LinReg
 y=a+bx
 a=113.2310847
 b=4.0857091E-4
```

Now, in the words of the chain gang boss in the movie, Cool Hand Luke, "What we've got here……is a failure to communicate." That is to say, we have been singularly unlucky. Not only do we have information we didn't ask for, we don't have Pearson's *r*, which we did ask for. What has gone wrong here is not the extra information; it's the missing information. For reasons unknown the TI-83 calculator right out of the box does not present *r* without a little coaxing. That coaxing is of the following keystroke form:

**2nd > CATALOG…**

At this time you may marvel at the relatively short list of choices you were presented with in the **Stat > Calc** sequence you did above. Compared to that, the list we have now is seriously long…

```
CATALOG
►abs(
 and
 angle(
 ANOVA(
 Ans
  (etc.)
```

Arrow down, down, down, until you get to the D's, and execute this keystroke sequence:

**DiagnosticOn > Enter > Enter**

(Yes, we mean **Enter** two times)

Fortunately this **DiagnosticOn** set of keystrokes only has to be done once. The **DiagnosticOn** tells the calculator that Yes, you want to see Pearson's *r*.

Now let's start at the top…

**Stat** > **Calc** > **`8:LinReg(a+bx)`** **>** $L_5$, $L_6$ **> `Enter`**

Here's what we see on our calculator at this point:

```
8:LinReg(a+bx)  L5, L6
```

and…

```
LinReg
 y=a+bx
 a=113.2310847
 b=4.0857091E-4
 r²=1.817941E-6
 r=.0013483102
```

We have succeeded in getting Pearson's correlation. All those decimals just show the calculator's sense of humor; we would most likely just go with $r = 0.001$.

## 5.2 The regression line

**Example 5.6: Defibrillator Shock and Heart Attack Survival Rate**

Studies have shown that people who suffer sudden cardiac arrest (SCA) have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered? Here is the data from this example: recall that $y$ = survival rate (percent) and $x$ = mean call-to-shock time (minutes.) These data are from a cardiac rehabilitation center (where cardiac arrests occurred while victims were hospitalized and so the call-to-shock times tend to be short) and for four communities of different sizes.

| | | | | | |
|---|---|---|---|---|---|
| mean call-to-shock time, x: | 2 | 6 | 7 | 9 | 12 |
| survival rate, y: | 90 | 45 | 30 | 5 | 2 |

As you did with the foal data of Example 5.6, enter these pairs into your calculator. We will again use List5 and List6 as we work through the problem on the TI. After the data is entered, duplicate what we did earlier, except that you do not have to go through all of that **DiagnosticOn** stuff – the calculator will stay in the **On** mode until you change it. (So don't change it!)

**Stat** > **Calc** > **8:LinReg(a+bx)** **>** $L_5$, $L_6$ **> Enter**

Here's what I see on my calculator now:

```
8:LinReg(a+bx)  L5, L6
```

and…

```
LinReg
 y=a+bx
 a=101.3284672
 b=-9.295620438
 r²=.921745124
 r=.-.9600755824
```

Do not worry that the answer here does not agree exactly with the "by-hand" solution in POD; the differences are due to round-off errors. Calculators, bless their hearts, have a great deal more patience for 10-digit arithmetic than the garden variety human.

At this point you have performed the regression and have the best fit line in hand. Reading the calculator screen gives us: $\hat{y} = 101.33 - 9.30x$ . We can also make the scatter plot as discussed in section 3.3. Before we go further you should set up the scatter plot as discussed in section 3.3 – we are going to plot the best fit line on that scatter plot, but we need to get it set up for the scatter plot first. When the scatter plot and Window is set to your satisfaction, please continue.

OK, now that we're happy with the scatter plot, let's retrace our steps. Return to the sequence of keystrokes that look like…

**Stat** > **Calc** > **8:LinReg(a+bx)** **>** $L_5$, $L_6$ **> Enter**

We're going to alter this sequence slightly, so that (eventually.) it looks like this:

**Stat** > **Calc** > **8:LinReg(a+bx)** **>** $L_5$, $L_6$, $Y_1$ **> Enter** .

It appears pretty simple but the keystrokes to get that $Y_1$ will be a bit convoluted so please bear with us. What we're going to do is "save" the least squares regression line and "paste" it into the calculator's graphing window. If you have already graphed functions with the TI-83, you know how to use the "**Y=**" key to set up a function definition. We will be getting the same result, a function, but it will be entered for us by the calculator after it does the linear regression calculations. Got it? Here we go… Enter this sequence of keys, and pause after entering the lists your data are stored in:

**Stat** > **Calc** > **8:LinReg(a+bx)** **>** $L_5$, $L_6$

Now we'll add some keystrokes, starting with a comma…

**,** > **VARS** > **Y-VARS** > **1:Function** **>** $Y_1$ **>** **Enter…**

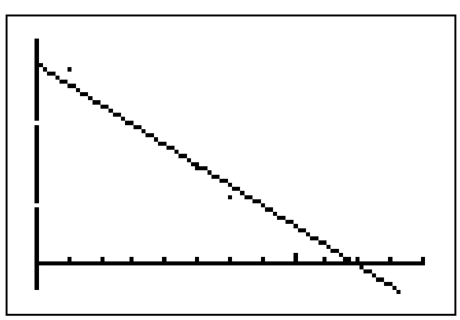
Now wait! We're not done yet. You should now see the following on the screen

```
LinReg(a+bx) L5,
L6, Y1
```

with a blinking cursor after the $Y_1$ **.** Now, press **Enter** one more time…

```
LinReg
 y=a+bx
 a=101.3284672
 b=-9.295620438
 r²=.921745124
 r=.-.9600755824
```

Now go back and graph the scatter plot -- there should be a new kid on the block – er, screen:



## 5.3 The residual plot

We will now assess the plausibility of the straight-line model using the residual plot. Recall that we are basically looking to see if there is any indication of the pattern of plots deviating from the straight line our calculations give us. If we see some curvature, for example, we will be suspicious that the straight line we used as a model for the relationship between x and y might have been simpler than reality demands.

**Example 5.10: Tennis Elbow**

One factor in the development of tennis elbow is the impact-induced vibration of the racket and arm at ball contact. Tennis elbow is thought to be related to various properties of the tennis racket used. The accompanying data are measurements on $x$ = racket resonance frequency (Hz) and $y$ = sum of peak-to-peak accelerations (a characteristic of arm vibration in m/sec/sec) for $n$ = 14 different rackets.

| Racket | Resonance (x) | Acceleration (y) |
|---|---|---|
| 1 | 105 | 36.0 |
| 2 | 106 | 35.0 |
| 3 | 110 | 34.5 |
| 4 | 111 | 36.8 |
| 5 | 112 | 37.0 |
| 6 | 113 | 34.0 |
| 7 | 113 | 34.2 |
| 8 | 114 | 33.8 |
| 9 | 114 | 35.0 |
| 10 | 119 | 35.0 |
| 11 | 120 | 33.6 |
| 12 | 121 | 34.2 |
| 13 | 126 | 36.2 |
| 14 | 189 | 30.0 |

Once again, enter these data into your calculator. We will use List1 and List2. Once you have the data entered, duplicate your efforts to get the scatter plot and the best-fit line for these data on the screen. To get the residual plot we will proceed from that point. So that we are on the same playing field, what we see on our screen after we have done the regression is shown at right.

```
LinReg
Y=a+bx
a=42.37454497
b=-.064520998
r²=.6006256099
r=-.7750003935
```

We know that for simple regression, a residual plot with $\hat{y}$ on the horizontal axis will have the same shape as a residual plot with $x$ on the horizontal axis. It is easier and takes fewer steps on the calculator to get a residual plot with $x$ on the horizontal axis, so this is our recommended procedure. By now you are familiar enough with the TI-83 to know that it performs some statistical calculations just in case you need them; it may not surprise you to learn that the TI-83 has already calculated the residuals and is waiting patiently for you to do a residual plot. In fact, the TI-83 calculates the residuals each time you perform the `LinReg` procedure and stores them for your use. The only problem with this automatic calculation is that you have to remember to manually store the residuals if you don't want to lose them while doing a different regression, such as re-doing a regression calculation after deleting an influential point.

Creating a residual plot is easy, once you remember that a residual plot is really nothing more than a scatter plot of residuals vs. $x$ (or, if you wish, $\hat{y}$) variables. Thus, we need to refresh our memories about how to get scatter plots. Remember? Back in section 3.3? Those Vermont sugarbushes? OK, here we go – let's first set up the plot for a scatter plot with this familiar sequence:

**2$^{nd}$ > STAT PLOT > `Plot1`**

We use Plot1 here, but you may, of course, use whichever Plot you wish. The "`Type:`" is scatter plot, the upper left choice, and you can pick your favorite "`Mark`." Now we get to the part that is new about the residual plot – just where are the residuals??? What do we choose for our `XList` and `YList` values? Since we are using the $x$ rather than $\hat{y}$ for our horizontal axis, XList is whichever list contains the $x$ values – in our case, $\text{List}_1$. `YList` will contain the residuals, wherever they are, and as it turns out they are in a special list called `RESID`. This list is maintained by the TI-83 and as we mentioned, updated each time we do a regression. We don't access the RESID list through the Edit screen, but by a separate set of keystrokes. Place your cursor on the `YList` line in the Plot Choice menu, and key the following strokes:

**2$^{nd}$ > LIST**

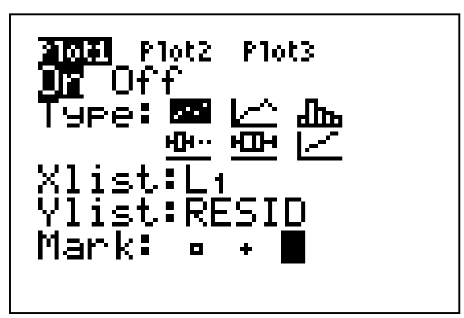
```
NAMES OPS MATH
1:RESID
2:Y
```

and you should see a screen more or less like this one.

The reason that the screen will be "more or less" like the one shown at right is that with calculator use, data files are sometimes saved as lists. If you are borrowing someone else's calculator they may have already created and named some data files. To select `RESID` as your list of choice, the keystrokes should be:

**`NAMES` > ▼▼▼…▼ > `RESID` > ENTER**

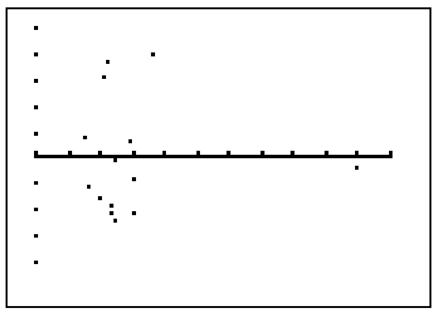
(That is, you will have to arrow down the alphabetical list until you get to `RESID` and then Enter. The calculator will then place the list of residuals in the Plot1 screen. Your Plot1 screen should now look like this (unless you chose different Lists for your data).



Exit this screen, and **ZoomStat** to see your scatter plot. After choosing our **WINDOW** as…

```
WINDOW
Xmin=90
Xmax=200
Xscl=10
Ymin=-2
Ymax=2.5
Yscl=.5
Xres=1
```

…and choosing **GRAPH** we get the residual plot:



With the residual plot in hand, we can assess the plausibility of the straight line model.

## 5.4 Conclusion

In this chapter our capability to use the graphing features of the TI have been greatly enhanced. We have practiced with the Edit screen, the Plot choice screen, etc. and added regression techniques and residual plotting to our growing list of TI tools. We will use all these skills again in Chapter 13.