

Assignment 1. Introduction to Spreadsheet and Regression Analysis

© 2010 Samuel L. Baker

This script for assignment 1 will guide you through these steps:






1. Getting the assignment's data
2. Graphing the data and drawing a line through the points by hand
3. Calling up the Excel spreadsheet program
4. Entering the data into the spreadsheet
5. Calculating the slope and intercept of the least squares regression line through the data





The list of what you need to turn in for assignment 1 is on page 21.


The instructions in this handout are for Microsoft Excel.

If you are in the lab, **be sure you have your** storage device – a **flash drive** is best – for saving your work.

Characters you are supposed to type use this monospace typeface.

Other keyboard keys are indicated like this:      .

The arrow keys are indicated like this:     . If you press an arrow key and get a number instead, press the  key once to turn Number Lock off.

+V means to press and hold the Ctrl key and then press V or v. Don't press the + key.


Get the data

To get the data, browse to the assignment web page, at <http://hspm.sph.sc.edu/Courses/J716/Data1.html>

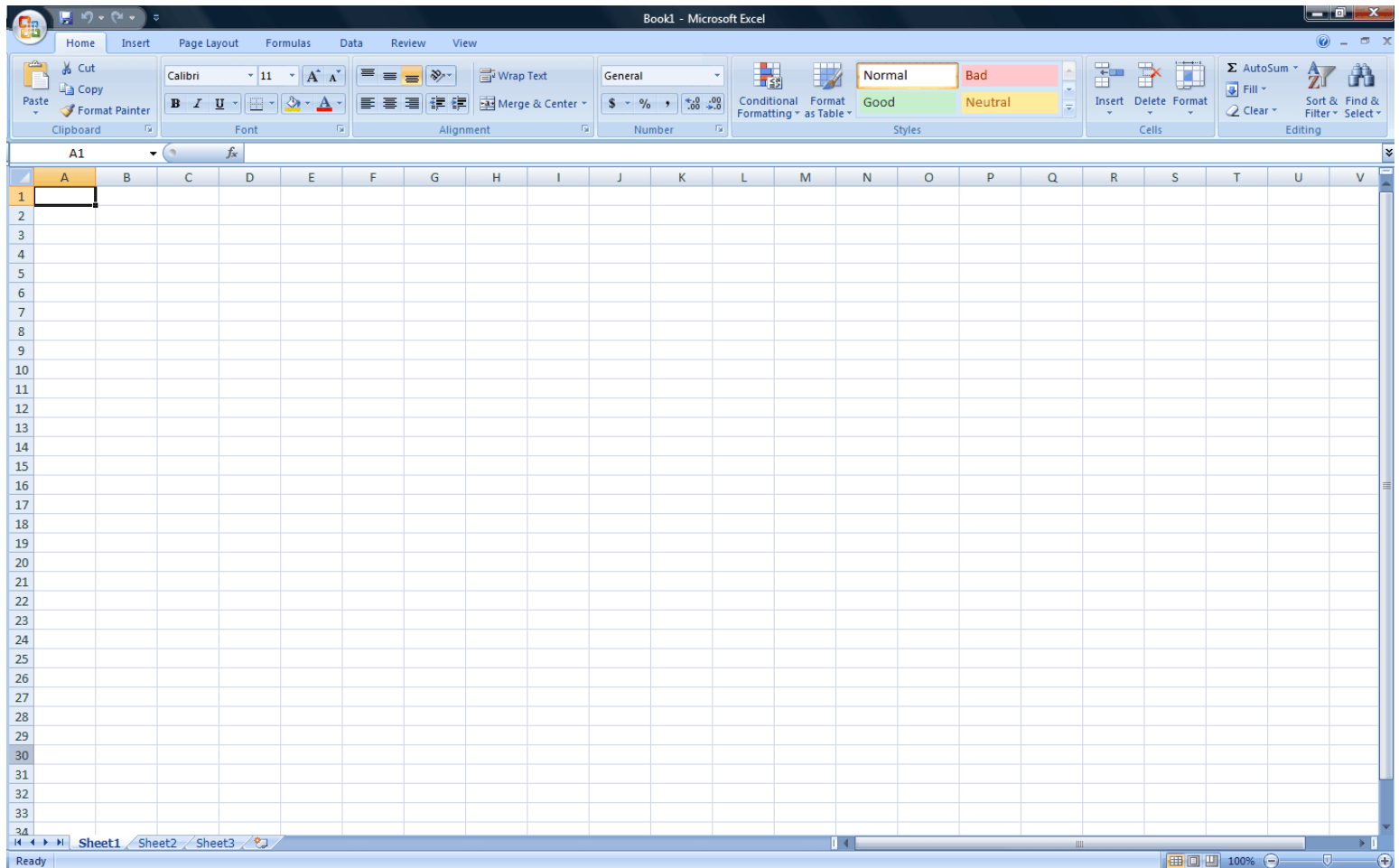
The web page has instructions for copying and pasting the data directly into a spreadsheet. You can try those instructions, or you can hand-type the data numbers into a spreadsheet, using the method described below. Either way, your next step is to start up your spreadsheet.

Starting the spreadsheet

Start up Excel (or the spreadsheet of your choice). On a Windows computer you can click the Start button. Move the mouse up to All Programs in Windows Vista or Programs in Windows XP. Allow the menu of programs to appear. Move the mouse to Microsoft Excel and click on it. If Microsoft Excel does not show on the menu, look for the Microsoft Office folder and click it.

Once Excel is running, I like to maximize both its windows (the inner one and the outer one), if they aren't maximized already. In Windows, maximize each window by clicking on the  near the window's upper right corners.

Excel 2007 looks like this when first opened and fully maximized:



This is your electronic accountant's paper. It has rows, designated by numbers, and columns, designated by letters.

Each intersection between a row and a column is a cell, designated by its letter and number. For example, the upper left cell is A1.

You have many rows and columns available -- many more than will show on your screen at once.

You also have multiple notebook pages, designated by tabs like `_Sheet_1_/`. (Quattro Pro designates the tabs with letters, like `_A_/`.) Each notebook page is like a whole spreadsheet, with its own rows and columns.

One cell is highlighted by the "cell selector." At the start, it's cell A1. If you type something, what you type will appear in the selected cell. The arrow keys, as well as `(Home)`, `(PgUp)`, etc., move the cell selector around the spreadsheet. Try using them to move around now. You can also use a mouse to move around by left-clicking on any cell.

Entering the data

Our task is to put the data from the web page into columns B and C as shown below. (Leave column A blank for now.) Your data in column C will be different from what is shown here, because the web page gives different numbers every time it is opened..

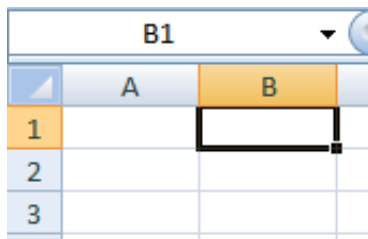
| | A | B | C |
|---|---|-----|----|
| 1 | | X | Y |
| 2 | | 100 | 40 |
| 3 | | 200 | 45 |
| 4 | | 300 | 50 |
| 5 | | 400 | 65 |
| 6 | | 500 | 70 |
| 7 | | 600 | 70 |
| 8 | | 700 | 80 |

If you copied the data using the technique described on the web page, you can click on cell B1 and paste it in. If that works, you can skip ahead to the **Drawing the Graph** section. If you prefer to type the numbers in by hand, which is a good idea if you are new spreadsheets, here are instructions:

What you do

Start with a blank spreadsheet. Freshly start Excel, or click the round button in the upper left corner, then New, then Blank Workbook. (In Excel for the Mac or Excel 2003, use File New from the menu to get a new blank spreadsheet.)

Use the arrow keys or the mouse to move the cell selector to cell B1, the first cell of the B column.



Type:

X (just that one keystroke, for now)

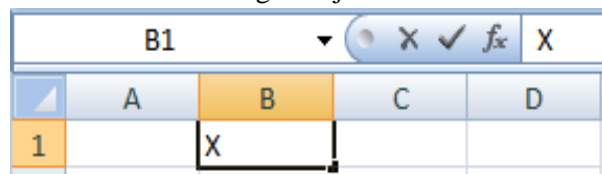
What it does

We don't want any clutter as we are getting started.

We reserve the A column for labels to tell us what the rows mean. Put the X numbers in the B column.

“X” is the header for this B column of data.

As you type letters or numbers, what you type appears on the screen in two places. One is the cell itself. The other is an editing area just above the column headers.



You can use either place to change what you are typing.

If you want to change something while you are in the middle of entering something into a cell, don't use the arrow keys. They will move the cell selector to a different cell. Instead, do one of these things:

- Use **←Bksp** to rub out your entry, character-by-character, or
- **Esc** will start the cell entry over from scratch, or
- Click with your mouse in the cell or in the editing area.
-

To change what is in a cell once you have moved on to a different cell, move the cell selector back to the cell. Type the new entry, which will replace the old entry, or press **F2** or double-click the cell to edit the old entry, or click in the edit area above the lettered column headers.

To erase what's in a cell and leave it blank, move the cell selector there and press **Delete**.

Back to the keystrokes for this assignment. We have an X in cell B1 and your cursor is next to the X. Now:

Press **→** or click cell C1 with your mouse. This finishes the entry and moves right to cell C1.

Any arrow key finishes an entry (unless you've been editing, having pressed **F2**). The **Enter↵** key will also finish an entry, but moves the cell selector down. Clicking the mouse on a different cell finishes the entry and moves to the clicked cell.

Anyway, now we're in C1. Type:

Y **Enter↵** **←** Puts a "Y" in cell C1, and then moves the current cell to B2, underneath the X.

In the cells B1 and C1, the letters X and Y appear on the left edge of their columns. Excel left-justifies words and right-justifies numbers, unless you say otherwise.

We should now be in B2. Move the cell selector there if it is not there already. Type:

100 **Enter↵**
 (If the cell selector does not move down when you press **Enter↵**, use the down **↓** arrow key instead.) Enters the values for X in a column going down, starting with B2.
 200 **Enter↵**
 300 **Enter↵**
 400 **Enter↵**
 500 **Enter↵**
 600 **Enter↵**
 700 **Enter↵**

Tip: The numeric keypad can help you enter numbers quickly, if Num Lock is on. Pressing **NumLock** toggles Num Lock on and off. The keypad's Enter key works just like the **Enter↵** key.

Move up and over to cell C2, using arrow keys or a mouse click.

Moves to the cell under the “Y” label.

Type in each of your Y values in turn. Press or after each one to move down to the next cell in column C.

Fills in the Y column.

With that, we’re done entering the data.

Is the spreadsheet adding characters that you didn’t type? Sometimes spreadsheets try to help you with cell entries by suggesting what should go in the cell. They do that by adding characters in reverse video (dark background with white letters) to the right of where you are typing. If that happens, ignore the reverse video characters and keep typing. They will go away as you type. If there are still some reverse video characters left when you finish typing, them.

It’s now time to draw your graph. If you have run out of time and want to quit the spreadsheet program, please turn ahead to page 18 for details on how to save and quit, so that you can pick up later where you left off.

Drawing the graph

Please draw the graph before you have the computer calculate the least squares regression line. We don’t want the computer to influence your judgement.

Find some graph paper. I’ll wait.

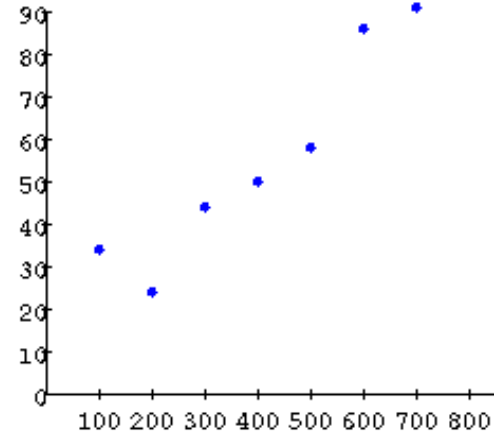
If you can’t find any graph paper, use the next page.

The image shows a full page of graph paper. It features a grid of 10 major columns and 10 major rows. Each major cell is further divided into a 2x2 sub-grid, resulting in a total of 400 small squares. The grid lines are thin and black, set against a white background. The grid covers the majority of the page, leaving margins at the top, bottom, and sides.

Plan your graph so the X axis goes from 0 to at least 800 and the Y axis goes from 0 to 100.

Your graph should look like this, except that your data points will be different.

Graph each of your seven data points. For example, if your first data point (X,Y) is (100, 35), your first dot should be about where the first dot is in the diagram to the right.

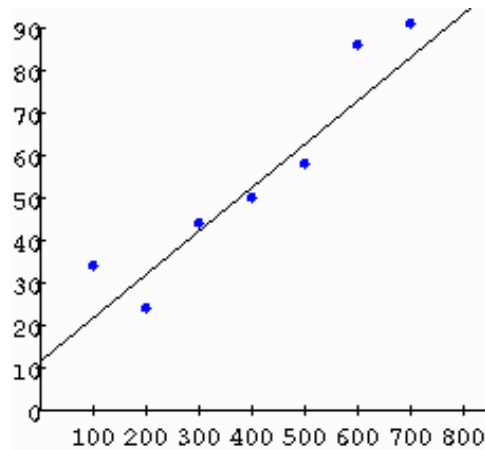


Once all seven points are marked, draw what looks to you like the best straight trend line through the points. We will call this line the “eyeball line.” Your graph should look something like the diagram below.

Extend the eyeball line to where X is 800. Read off your graph the predicted value for Y for X=800. In the diagram to the left, the line is about 92 over the place where X is 800. (I should have made the Y axis go up to 100. Then I’d have a better idea what the Y value is.

Write down the intercept of your eyeball line. This is where the eyeball line hits the Y-axis. It’s about 12 in this diagram.

A best-fit line I drew by eye. You can do better.



Calculate the slope of the eyeball line. To do this, use the two Y values you just wrote down. Let us designate your Y value at X=800 as Y_{800} , and your intercept as Y_0 . The slope is the difference between the Y values ($Y_{800} - Y_0$) divided by the difference between the X values (800-0, which is just 800).

You can now write out the formula for your eyeball line: $Y = \text{intercept} + \text{slope} \times X$. In terms of Y_0 and Y_{800} , the formula for the eyeball line is:

$$Y = Y_0 + \frac{Y_{800} - Y_0}{800} X$$

For example, in the graph above, Y_0 is 12. Y_{800} is 92. $Y_{800} - Y_0$ is 80. Divide that by 800 and you get 0.1. The equation for that eyeball line is $Y = 12 + 0.1X$

When your graph is ready and the slope and intercept of your line are calculated, go back to your spreadsheet.

A spreadsheet that can calculate the least squares line

At the right is the formula we will use for the slope of the least squares line through a set of points. It is equivalent to the one in SIMPLE REGRESSION THEORY, but slightly simpler to calculate. This formula may look intimidating, but we can implement this in our spreadsheet if we take the job one task at a time. We make the formula into a recipe:

$$\hat{\beta} = \frac{\sum_{i=1}^N (X_i - \bar{X})Y_i}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

1. We must calculate \bar{X} , the mean of the X values.
2. The $(X_i - \bar{X})$ means we must calculate how much each X value deviates from the mean of X.
3. We must multiply each of those X deviations by its corresponding Y value.
4. We must then add up the products of those multiplications. That will give us the numerator.
5. We must multiply each X deviation by itself.
6. We must add up the products of those multiplications. That will give us the denominator.
7. We have to divide the numerator by the denominator.

Step 1 is to get the mean of the X's. We'll get the mean of the Y's, too, because it's not much more work, and because we will need that to calculate the intercept.

Here is what the spreadsheet will look like when we have the means calculated:

| J | A | B | C | D |
|----|---------|------|------|------|
| 1 | X | Y | Xdev | |
| 2 | | 100 | 40 | -300 |
| 3 | | 200 | 45 | -200 |
| 4 | | 300 | 50 | -100 |
| 5 | | 400 | 65 | 0 |
| 6 | | 500 | 70 | 100 |
| 7 | | 600 | 70 | 200 |
| 8 | | 700 | 80 | 300 |
| 9 | | | | |
| 10 | Sum | 2800 | 420 | |
| 11 | Average | 400 | 60 | |

In this spreadsheet, the Sum row has sums of the entries in the columns above. The Average row has the means of the entries in the columns above. The column headed "Xdev" has the deviation of each X value from the mean of the X's. Each entry in that column is $X_i - \bar{X}$ for a particular i value. For example, the mean of the X's is 400, and the first X value is 100, so the first Xdev, $X_1 - \bar{X}$, is $100 - 400 = -300$. That's the value in cell D3.

What you type

Move the cell selector down and over to A10.

What it does

Skipping a line between our data and our sums looks good, and gives us a space that we will use later.

Type:

Sum

Puts in labels for what will be the rows of sums and averages.

Average

Move the cell selector to B10. Then type:

=sum (B2 : B8)

This puts a formula in B10 that sums all the X values above it.

or

=sum (b2 : b8)

=sum is a function that adds up the contents of the cells indicated inside the parentheses.

Formulas are what make electronic spreadsheets really useful. A spreadsheet can calculate formulas involving many different mathematical functions, saving you the work of doing calculations by hand.

In our sum formula, the leading equals = sign starts a formula in Excel, telling it that “sum” is meant to be a function name, not a word that we want to see in the cell. (In Quattro Pro, use @sum in place of =sum. @sum works in Excel, too.)

The B2:B8, inside the parentheses, is the “block” or “range” of cells you are summing. The : separates the two ends of the block. (Quattro Pro uses .. to separate ends of blocks. The .. separator also works in Excel.) A block can be vertical, like B2:B8. It can be horizontal, like B2:E2. It can also be rectangular, like B2:E8. To designate a block, you name the cell at one end, put in the colon or the two dots, and then name the cell at the other end or opposite corner.

When you press Enter, you should see 2800 in cell B10. If you made a mistake, you might instead see ...

#NAME? One way Excel tells you that you have made an error is to put #NAME? in the cell. This tells you that your function name is wrong, or that you left something out, such as the first parenthesis, perhaps. Excel can detect some errors. If you leave out a closing parenthesis, it might put one in for you. By the way, when you type a formula, UPPER CASE, lower case, or mIxEd CaSe doesn't matter.

Now for the average (mean) formula:

Move the cell selector to B11. Type:

=average (b2 : b8)

Averages the X values.

In Quattro Pro or Lotus 1-2-3, type
@avg(b2..b8)

You should see the mean of the X's, 400, in B11. The =average function divides the sum of the numerical values in the block by the number of non-blank cells in the block. That means, for example, that we must not include B1 in the block, because B1 is not blank. =average would divide the sum by 8 instead of 7.

Now let's do some spreadsheet magic. Rather than retype the sum and average formulas in C10 and C11,

we can copy the formulas in B10 and B11 to C10 and C11. Here is how:

First, as a general rule, Windows programs' commands go object-action. First, you specify what object you want to work on. Then you specify what action you want to take with it.

Our first intended object is the block of cells B10:B11, the cells we want to copy from. We want to “select” this block.

To **select B10:B11**, we can do one of two things:

1. Use the keyboard like this:
 - a. Move the cell selector to B10
 - b. Press and hold ⇧ Shift
 - c. Press ↓ or left-click with your mouse on B11.

Or

2. Use the mouse:
 - a. Move the mouse pointer to B10.
 - b. Press and hold the left mouse button, and move (“drag”) the mouse to B11.
 - c. Release the mouse button.

| | A | B | C |
|----|---------|------|----|
| 1 | | X | Y |
| 2 | | 100 | 40 |
| 3 | | 200 | 45 |
| 4 | | 300 | 50 |
| 5 | | 400 | 65 |
| 6 | | 500 | 70 |
| 7 | | 600 | 70 |
| 8 | | 700 | 80 |
| 9 | | | |
| 10 | Sum | 2800 | |
| 11 | Average | 400 | |

Either way, the selected block shows a heavy outline.

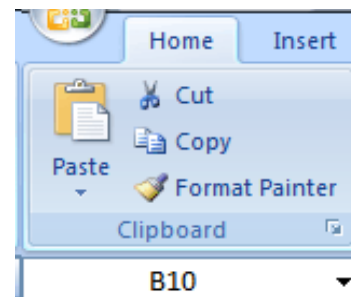
In older Excel versions, you may have to move quickly when you use the mouse method to select a block. If you click and hold the mouse on a cell, wait a couple of seconds, and then drag, the Excel 2003 may assume that you want to move the cell, rather than extending the selected block.

When you select a block, you can start at the cell at either end of the block. In this example, you could have clicked first on B11 and then extended the block to B10.

Once the block B10:B11 is selected as our object, we carry out the action, which is to **copy that block to the clipboard**. The clipboard is a feature of your computer’s operating system, be it Windows, Mac, or other. The clipboard works in the background as a holder for whatever you want to copy or move. You can’t see the clipboard, but it’s there.

There are three ways to copy the selected cells to the clipboard:

1. Click where it says Copy, under the Home tab at the upper left. (Click the Home tab if you don’t see this.)
- or
2. Keyboard shortcut: Press Ctrl+C (This means hold down Ctrl and press C.).



Whichever of these you do, the border of the selected block changes to moving dots, and a copy of B10:B11 goes to the clipboard. The clipboard copy has not just the numbers you see, but also the underlying formula information.

Now we make the copy. Object-action again. This time the object is the destination for the copy. Move the cell selector to C10 with the arrow keys or left-click on C10. That **selects C10** as the object.

Next is the action, which is **“pasting” from the clipboard to C10**. Notice that you have designated a single cell, C10, as the destination for the paste. The stuff in the clipboard will actually cover C10:C11, but all we tell the spreadsheet is where the pasted area starts. In general with spreadsheets, when you want to paste one copy of a block of cells, you tell it to paste to the upper left corner of the area the paste will occupy. If you designate a block of cells as the destination for pasting, you get multiple copies, one for each cell in the block. (We’ll use that technique soon.)

There are three commands you can use to paste the contents of the clipboard to the selected cell:

1. Click Paste
- or
2. Keyboard shortcut: Press **Ctrl+V**

The sum and average of the Y's should appear in column C.

A neat feature – that makes this work the way we want – is that cell references in formulas automatically adjust when copied. Here, the formula in B10 is `=sum(B2:B8)`. Its copy in C10 is `=sum(C2:C8)`. This is called “relative addressing.”

Now we're ready to fill in column D with the X deviations.

Move up to the top of column D, to cell D1.

Type:

xdev **Enter**

Puts “Xdev” in D1 and moves to D2.

In D2, type:

=b2-b11 (Wait! Don't press **Enter** or an arrow yet!)

Start to enter the formula, but don't finish yet, because we have to make a change.

The `=b2-b11` formula should be showing on your screen, with the cursor just after the final “1”. (If you did press **Enter** or an arrow, move the cell indicator back to this cell and press **F2** or click in the edit area above.)

Some explanation before we go on:

The = at the beginning of this formula is needed to tell Excel that this is a formula. Formulas have to start with one of these symbols: = + -

If you just type `b2-b11`, Excel will treat it as text.

Why this formula? Cell B2 has the X value. Cell B11 has the mean of the X's. `=b2-b11` is thus the first X value minus the X mean. That difference is Xdev for the first observation.

The reason I want to make a change to this formula is this: If you copy $-b_2 - b_{11}$ down the column, the relative addressing feature presents a problem. Going down the column, we do want the B2 to change to B3, B4, etc., to pick up each X value, but we don't want the B11 to change to B12, B13, etc. The B11 must stay the same in each copy, because we want to subtract the same number – the mean of the X's – from each different X value.

Here's how we tell Excel which cell references in the formula we want adjusted and which ones we don't. This is done by adding \$ signs to the formula. Specifically, we change B11 to \$B\$11.

In Windows Excel, you can do this:
Starting with the cursor after the "1" in the cell's formula, press:

F4 This changes the b11 to \$B\$11.

Each \$ means that what comes next is not changed as the formula is copied. \$B\$11 means that neither the column (B) nor the row (11) will be adjusted when this formula is copied. (The change from lower case "b" to upper case "B" doesn't matter.)

Pressing **F4** repeatedly would cycle through all four possible combinations of \$'s and cell references. Try it. Cycle around until you get back to \$B\$11.

Whether you typed \$'s directly or used **F4**, the formula should read $=B2 - \$B\11 .

The formula for the deviation of this X value from the mean of the X's is done.

When it does, press **Enter**.

Move the cell selector to D2 and copy it to the clipboard by one of the ways already discussed, clicking Copy or pressing **Ctrl+C**.

This copies D2 to the invisible clipboard so that you can paste the copy into other cells.

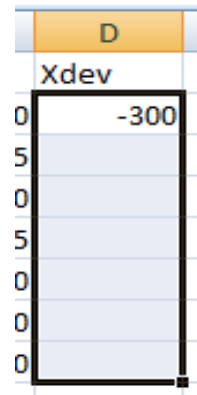
For the destination, select the block D2:D8.

You should see this:

You can:

1. Left-click on D2, hold down **Shift**, and press **↓** repeatedly.
or
2. Left-click D2, hold down **Shift**, and left-click on D8.
or
3. Move the mouse pointer to D2, press and hold the left mouse button and move the mouse to D8, and release the mouse button.

The selected cells are outlined. Moving dots may also be showing around the 300.



Paste to this block, with
 Paste or **Ctrl**+V

The destination block should fill in with numbers
 calculated from the formula you copied.

You can see the formula in any cell by clicking on it. The formula will appear in the window that is just below the icon bars near the top of the spreadsheet. Move up and down the D column and you can see how the formula from D2 copied. You should see that the first cell reference in each formula changes from B2, to B3, to B4, etc., as you go down the column, but that the reference to \$B\$11 stays the same.

To see all the formulas at once, use Show Formulas from the Formulas tab, or press **Ctrl**+`

The screenshot shows the Microsoft Excel interface with the 'Formulas' tab selected. The ribbon includes options like 'Insert Function', 'AutoSum', 'Recently Used', 'Function Library', 'Name Manager', 'Trace Precedents', 'Trace Dependents', 'Remove Arrows', 'Evaluate Formula', 'Watch Window', and 'Show Formulas'. The 'Show Formulas' button is highlighted. Below the ribbon, the spreadsheet is visible with columns A through F and rows 1 through 11. The D column contains the following data and formulas:

| | A | B | C | D | E | F |
|----|---------|-----------------|-----------------|-------------|---|---|
| 1 | | X | Y | Xdev | | |
| 2 | | 100 | 40 | =B2-\$B\$11 | | |
| 3 | | 200 | 45 | =B3-\$B\$11 | | |
| 4 | | 300 | 50 | =B4-\$B\$11 | | |
| 5 | | 400 | 65 | =B5-\$B\$11 | | |
| 6 | | 500 | 70 | =B6-\$B\$11 | | |
| 7 | | 600 | 70 | =B7-\$B\$11 | | |
| 8 | | 700 | 80 | =B8-\$B\$11 | | |
| 9 | | | | | | |
| 10 | Sum | =SUM(B2:B8) | =SUM(C2:C8) | | | |
| 11 | Average | =AVERAGE(B2:B8) | =AVERAGE(C2:C8) | | | |

A tooltip for the 'Show Formulas' button is visible on the right side of the spreadsheet, stating: 'Show Formulas (Ctrl+`)' and 'Display the formula in each cell instead of the resulting value. Press F1 for more help.'

On the Mac, Control+` does this. ` is on the same key as ~. In older Excel, choose Tools Options View Formulas from the menu.

With the formulas shown, we can see that, in the D column, the cell reference B2, with no \$ signs, adjusts as it goes down the column. \$B\$11, which has \$ signs, does not adjust.

Finishing the regression line calculation

Here's where we want to get: (Again, many of your numbers will differ from these because your Y values are different.)

| J | A | B | C | D | E | F |
|----|--------------|----------|------|--------|--------|----------|
| 1 | X | Y | XDev | XdevSq | Xdev*Y | |
| 2 | | 100 | 40 | -300 | 90000 | -12000 |
| 3 | | 200 | 45 | -200 | 40000 | -9000 |
| 4 | | 300 | 50 | -100 | 10000 | -5000 |
| 5 | | 400 | 65 | 0 | 0 | 0 |
| 6 | | 500 | 70 | 100 | 10000 | 7000 |
| 7 | | 600 | 70 | 200 | 40000 | 14000 |
| 8 | | 700 | 80 | 300 | 90000 | 24000 |
| 9 | | | | | | |
| 10 | Sum | 2800 | 420 | 0 | 280000 | 19000 |
| 11 | Average | 400 | 60 | 0 | 40000 | 2714.286 |
| 12 | Coefficients | | | | | |
| 13 | Slope | 0.067857 | | | | |
| 14 | Intercept | 32.85714 | | | | |

To get to this, we first need a column of squares of the X deviations in column E.

Move to E1. Type:

XdevSq

Labels the first cell of the E column.

The cell selector moves to E2. Type:

=d2*d2

This is the value in D2 multiplied by itself.

The * is the computer symbol for multiplication. The letter “x” is not used for this purpose.

Another way to square a number is =d2^2. The ^ is the computer symbol for raising something to a power. Either is OK here.

Let's copy this cell down the E column. We don't need \$ signs in this formula, because we want the d2's in this formula to be adjusted as we go down the column.

Left-click on E2, if you're not already there.

Copy that cell to the clipboard, using Copy or +C.

The formula for Xdev squared is copied to the clipboard.

With the cell selector in E2, hold down $\boxed{\leftarrow \text{Shift}}$ as you press $\boxed{\downarrow}$ six times.

Select the block where we want the formula copied, which is E2 down to E8.

or

Move the mouse pointer to E2, hold the left button down, and drag the mouse down to E8.

Paste, using Paste or $\boxed{\text{Ctrl}}+\text{V}$

The copy is made! Cells E2 through E8 fill in.

The last column we need to fill in is F. There we want the products of the X deviations and the Y values.

Go up to cell F1 and type:

Xdev*Y $\boxed{\text{Enter} \leftarrow}$

This looks like a formula, but it is text. There is no leading = sign. It serves as the column label.

In cell F2 type:

=d2*c2 $\boxed{\text{Enter} \leftarrow}$

C2 has the Y value. D2 has the X deviation. This formula multiplies them.

We now want to copy this cell down the F column, as we did with column E. We don't use \$ signs in the formula because we want both cell references adjusted as we go down.

Move the cell selector to F2, if it's not there already.

This copies cell F2, so we can paste it down the F column.

Copy this cell to the clipboard, using Copy or $\boxed{\text{Ctrl}}+\text{C}$.

Select the block F2 down to F8.

This selects the block where F2 will be copied.

As before, you can:

Hold down $\boxed{\leftarrow \text{Shift}}$ as you press $\boxed{\downarrow}$ six times.

or

Move the mouse pointer to F2, hold the left button down, and drag the mouse down to F8.

Paste from the clipboard, using Paste or $\boxed{\text{Ctrl}}+\text{V}$.

The F column is filled in with copies of F2, with the cell references adjusted.

For the least squares regression line's slope (the formula is to the right), we need the sum of the squares of the X deviations (XdevSq's) -- that's the denominator in the expression -- and the sum of the Xdev*Y's -- that's the numerator.

To get those sums, we will copy the sum formula we have in C10 and paste it across the 10th row. We will also copy the average in C11 and paste it across the 11th row, even though we don't really need it.

$$\hat{\beta} = \frac{\sum_{i=1}^N (X_i - \bar{X}) Y_i}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

Move the cell selector to C10, the cell containing the sum of the Y's.

Select cells C10:C11. (Use $\boxed{\leftarrow}$ Shift + $\boxed{\downarrow}$ or click and drag with the mouse.)

Get ready to copy our sum and average formulas from the C column to the other columns.

Copy this to the clipboard, with Copy or $\boxed{\text{Ctrl}}$ +C.

Notice that we are copying two cells at once.

Select the horizontal block C10:F10.

Paste from the clipboard. Use Paste or $\boxed{\text{Ctrl}}$ +V.

Presto! Columns D, E, and F are all added up and averaged. Easier than doing it by hand!

Rows 10 and 11 is both filled in, even though we pasted just to row 10. That is because we copied two cells, one beneath the other.

Verify that the formulas are correct: The Sum and Average of the Xdevs in column D should both be 0. That is because each Xdev is the difference between that row's X value and the mean of the X's. Add up all these differences and they should cancel each other out, giving a total of 0.

Did you get an “E” number? You may get a number like 5.7E-14 in D10. This is scientific notation, computer-style. E means 10 raised to the power of the number after the E. (This way of writing scientific notation goes back to FORTRAN, the first high level computer language.) 5.7E-14, for example, means 5.7×10^{-14} , which is 0.000000000000057.

Cumulative round-off error is what gives you numbers like 5.7E-14. A digital computer allocates a certain amount of memory space for each number it stores. Fractions that are not powers of 2, meaning anything other than 1/2, 1/4, etc., are rounded off after a certain number of decimal places. The difference between the what the number is supposed to be and how the computer stores it is called round-off error. When you add or multiply a bunch of numbers, each of which is slightly different from what it's supposed to be, your result will also be slightly different from what it's supposed to be. The round-off error cumulates. Hence the name “cumulative round-off error.”

If your number in D10 is something like 5.7E-14, you can regard it as a 0. However, if D10 has a number that is not tiny, there is an error in your formulas that you need to fix.

We're ready for the payoff, the regression line slope and intercept.

Move the cell selector to B12.

Type Coefficients

To remind us that these are estimates of the true parameters.

Move to A13 and type:

Slope $\boxed{\text{Enter} \leftarrow}$

Intercept $\boxed{\text{Enter} \leftarrow}$

The labels for the slope and intercept go in A13 and A14.

Move over to B13.

We'll put the slope formula here in B13.

Let's look at the slope formula again.

$$\hat{\beta} = \frac{\sum_{i=1}^N (X_i - \bar{X}) Y_i}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

The numerator of the fraction is the sum of Xdev*Y, which is in F10. The denominator is the sum of the XdevSq's, which is in E10.

The computer sign for division is / , so type:

+f10/e10

Makes the slope value ($\hat{\beta}$) appear in B13.

The selected cell is now B14.

The intercept formula goes here, below the slope formula.

The intercept formula is:

$$\text{Intercept} = (\text{Y mean}) - \text{Slope} \times (\text{X mean}).$$

In terms of cell references, this formula translates into:

$$\text{B14} = \text{C11} - \text{B13} * \text{B11}$$

In B14, type:

=c11-b13*b11

The intercept formula.

The spreadsheet is complete! We now have the slope and intercept of the least squares regression line through your data points.

To view your formulas (this is optional), press +` to see this:

| | A | B | C | D | E | F |
|----|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1 | | X | Y | Xdev | XdevSq | Xdev*Y |
| 2 | | 100 | 40 | =B2-\$B\$11 | =D2^2 | =D2*C2 |
| 3 | | 200 | 47 | =B3-\$B\$11 | =D3^2 | =D3*C3 |
| 4 | | 300 | 43 | =B4-\$B\$11 | =D4^2 | =D4*C4 |
| 5 | | 400 | 67 | =B5-\$B\$11 | =D5^2 | =D5*C5 |
| 6 | | 500 | 62 | =B6-\$B\$11 | =D6^2 | =D6*C6 |
| 7 | | 600 | 72 | =B7-\$B\$11 | =D7^2 | =D7*C7 |
| 8 | | 700 | 82 | =B8-\$B\$11 | =D8^2 | =D8*C8 |
| 9 | | | | | | |
| 10 | Sum | =SUM(B2:B8) | =SUM(C2:C8) | =SUM(D2:D8) | =SUM(E2:E8) | =SUM(F2:F8) |
| 11 | Average | =AVERAGE(B2:B8) | =AVERAGE(C2:C8) | =AVERAGE(D2:D8) | =AVERAGE(E2:E8) | =AVERAGE(F2:F8) |
| 12 | | Coefficients | | | | |
| 13 | Slope | =F10/E10 | | | | |
| 14 | Intercept | =C11-B13*B11 | | | | |

Saving your work

You now have created a template for doing simple regressions on any set of observations on X and Y. You can change the numbers in the X and Y columns, and the spreadsheet will instantly recalculate the regression line. Go ahead and change some numbers in the X and Y columns to see what happens. Then change them back to your original data before proceeding. **[Ctrl]+Z** is your Undo, if you make changes and cannot remember how it used to be.

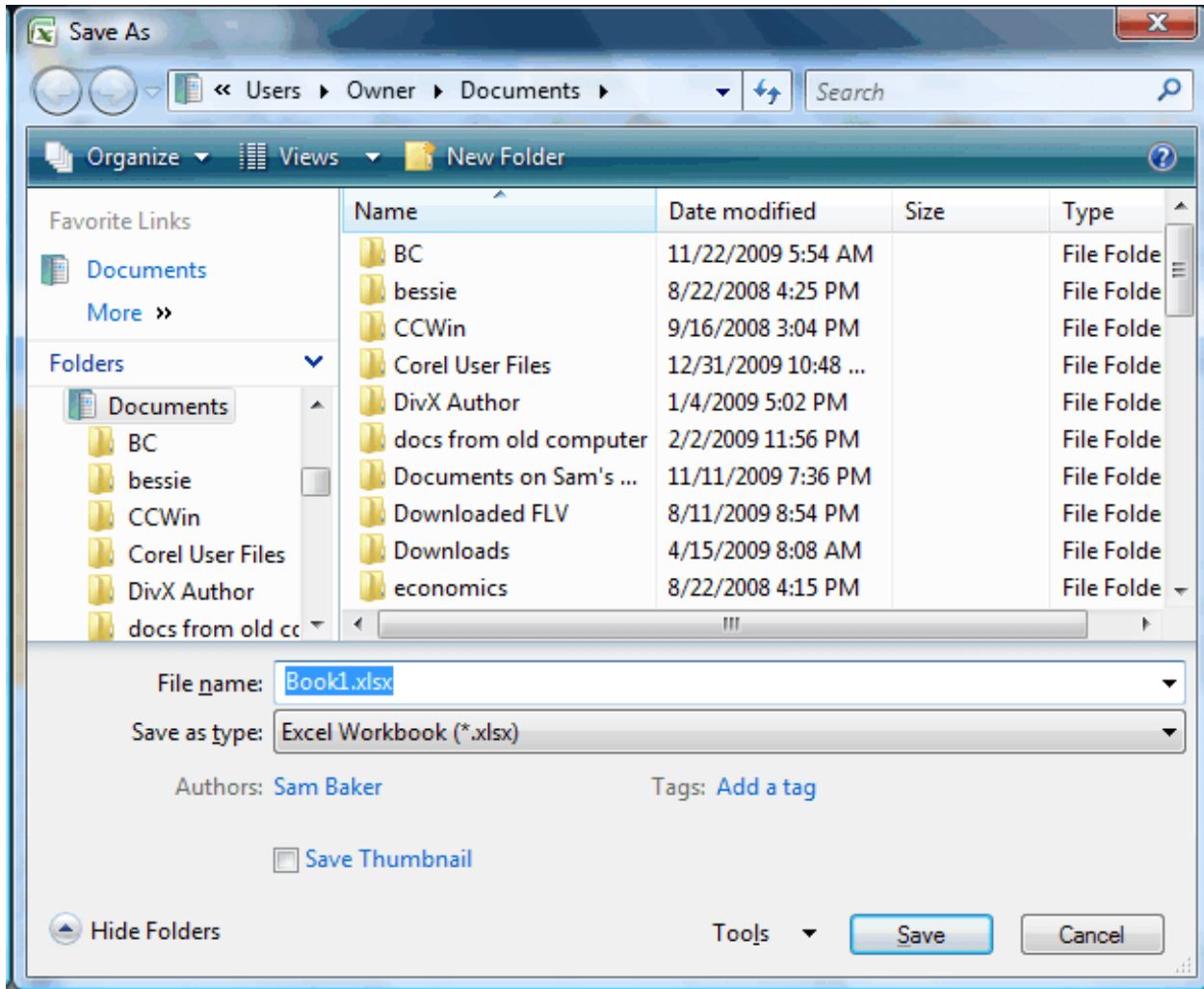
We'll be working with this spreadsheet template some more, so you need to save it.

To save your work:

1. If you are in the lab, slide your USB flash drive into one of the USB slots on the computer. Then
2. Click the round button, then Save (**F**ile, then **S**ave on other Excel versions) or press **[Ctrl]+S**

The Round button menu also offers a **Save As** choice. This is for saving something that you retrieved from a disk to a different file name. This will be useful later in the course, when you will be opening data files that I have prepared, working on them, and saving them to your storage device.

When you're saving a newly created file, the Save command works as if you had chosen Save As. You get a dialog box that asks for a file name. In the Lab, it looks like this:



Actually, this is how it looks on my computer at home. The lab computers' Documents folders have different stuff in them.

If you're in the Lab, you should save to your storage device. Click the slider in the left pane of the Save As dialog box. Move the slider until you see your flash drive. Click on that. Then, in the File name box, type the file name of your choice. Don't include any punctuation or any file type. Excel will supply that. For example, you can have Assignment1 in the File name box. Press . Excel will save the file on the device you selected, using the name Assignment1.xlsx. It adds the .xlsx ending for you.

If you're working on your own computer, you can save the file anywhere, of course. I recommend that you choose a name other than Book1.xls to avoid confusion later.

Printing your file

You may hand in a printout of your spreadsheet. If you are submitting your assignment on Blackboard, you can instead attach the spreadsheet file along with the word processor file that has your write-up.

Just in case you need the information, this is how to print:

Click the round button and then Print, or press $\text{Ctrl}+\text{P}$. You can see, and change if needed, which printer your computer will use. Print Preview will let you see what you will be printing, just to be sure there is no mistake.

Quitting and later resuming

To exit Excel, you may do any of these:

1. Round button and then the rectangular Exit button. (File Exit for other versions.)
2. Keyboard: $\text{Alt}+\text{F4}$
3. Mouse: Click on the X in the upper right corner of the screen.

To resume your work later, start up Excel as you did at the beginning of this assignment. Then use one of these to “open” your file:

1. Round button, Open. (File Open for other versions)
- or
2. Keyboard ($\text{Ctrl}+\text{O}$)

In the Open File dialog box, use the left pane to find your storage device. Click on the place where you stored your file. Then find the file name or type the name of the file you want in the File name box.

Completing this assignment

The equation of your regression line is $Y = \text{Intercept} + \text{Slope} \times X$. For example, my results were
 Slope 0.067857
 Intercept 32.85714 so my equation would be $Y = 32.85714 + .067857X$.

(If your slope and intercept came out the same as mine shown here, most likely you copied Y data from this book. Please use the data web page to get your Y numbers.)

This concept -- being able to write an equation with the slope and intercept in the right places -- is crucial. Fortunately, it's not too hard. Just be sure you get it.

To complete this assignment, draw your calculated regression line on your graph. To do this, plug two values of X into your equation, calculating the corresponding Y value for each. (I suggest using 0 and 800. The Y value for $X=0$ is just your intercept: If your equation is $Y = 32.85714 + .067857X$, plugging in 0 for X gives you $Y = Y = 32.85714 + 0 = 32.85714$. For the other X value, you might as well use 800, because below I ask you for the predicted value of Y if X is 800. If X is 800 and $Y = 32.85714 + .067857X$, then $Y = 32.85714 + .067857 \times 800 = 87$, approximately.) Plot both X,Y points on your graph. Draw the line through the two. How close are the regression line and the line you drew?

Write a sentence or two discussing the pros and cons of each regression method, eyeball and least squares. Which method do you think is better? Why? (We'll discuss the theory behind this in the next class, but I'd like you to think about it first.)

Use your least squares regression equation to calculate a predicted value for Y if X is 800. (Multiply 800 times the slope and add the intercept.) If you have not done so already, get your eyeball line's predicted

value for Y if X is 800. (Extend your eyeball line over to where $X=800$. Read the Y value off the graph.)

Turn in to me, before or at the beginning of class:

- a. The graph with the data points and **both** lines on it. (If you are submitting your assignment by e-mail, bring your graph to class or send it by regular mail, unless you have a scanner so that you can create an image of your graph.)
- b. Your prediction of Y for $X=800$, based on your eyeball line.
- c. The slope and intercept of your eyeball line, as best you can calculate it.
- d. The printout of your spreadsheet. (Or a copy of your spreadsheet file, if you are submitting by Blackboard.)
- e. The slope and intercept of the least squares line.
- f. The predicted Y value for $X=800$ based on your least squares line.
- g. The discussion of the pros and cons of the eyeball and least squares methods.

If you get this done before noon on the day of class, please send me an e-mail with the slopes and intercepts of your eyeball and least squares lines. We'll use the information during the class.

If you want to submit your write-up, your spreadsheet file, and maybe a scan of your graph, all on Blackboard, here is how: Do not hit Submit until you have added all your files. Attach your first file by browsing for it. Then hit the "Add another file document" button. That should let you browse for your second file. Repeat until you have added all the files you want. Hitting Save along the way is useful if you do not have all your files ready to attach right now. When you can see a complete list of the files you had wanted to attach, then and only then hit Submit. Once you hit Submit, Blackboard will not allow you to go back and attach or change anything.

It is OK to submit some of your work on paper and some on Blackboard.

How are you doing? If you have the time and energy, go right on and do assignment 1A for this week. Read Simple Regression Theory II and then do the assignment. If getting started with the spreadsheet was tough, do assignment 1A along with assignment 2 for next week.