Assignment 2. Predictions, Graphs, and Heuristic Inferences from Graphs

© 2006 Samuel L. Baker

The least squares method of drawing a line through data gives good results if the assumptions discussed in Simple Regression Theory II are true. The least squares method cannot detect whether or not those assumptions are true. You have to do additional tests to get an idea.

The additional test we will use in this assignment is a graph of the data. We will visually inspect the graph to see if the assumptions behind using least squares apply.

Our job is to project December's utilization for three clinics, based on the trend of the preceding eleven months. Here are the data:

		Clinic visits		
	Month	No. 1	No. 2	No. 3
January	1	426	310	539
February	2	568	474	573
March	3	724	613	608
April	4	482	726	642
May	5	695	814	677
June	6	881	877	711
July	7	804	914	746
August	8	833	926	781
September	9	1084	913	815
October	10	758	874	1274
November	11	996	810	884

We will use our regression template from assignment 1A to calculate a least squares line for each of these clinics. We'll use those lines to make a prediction for each clinic. And we'll make graphs of each clinic's data. After looking at the graphs, we'll decide if we like the least square prediction for each clinic.

Our first task is to open the spreadsheet file in which you saved Assignment 1A. There are two ways to do this:

Alternative 1: Opening a file using My Computer

This method conforms best to the object-action way of doing things in Windows.

What you do	**What it does**
Minimize any open windows. Double-click the My Computer icon on the Windows desktop.	Shows all your available disk drives.
Double-click on the drive where you saved Assignment 1A. In the lab, this should be Drive A: or Drive E:	Shows all the files in the root directory of your drive.

Double-click on your spreadsheet file. Its icon should represent Excel or the spreadsheet you used to create it.

Excel (or your spreadsheet if different) will start up (or pop up, if Excel was already running minimized) and load the file. This can take a few seconds.

Alternative 2: Open the file from within Excel

Click the Start button in the lower left corner of the screen.
Move the mouse pointer to Programs.
Move over and down to Microsoft Excel and click.

Starts the Excel program.

Select File then Open from the top menu,
or press Ctrl+O
or click on  .

Brings up a file open dialog box so you can chose which file to open.

If you saved your assignment to a diskette, type A: Enter↵ in the File name box.

Shows a list of the spreadsheet files on your diskette.

Double-click on the file that has your Assignment 1A spreadsheet.

Loads the file into Excel.

Adding Blank Rows to Accommodate the Data

In the Assignment 1 and 1A we had seven rows of data. This time we have eleven, so we must add four rows to the data part of the spreadsheet.

Move the cell selector to A3. We'll add the rows there. (Actually, you can go to any cell between rows 2 and 8, but the directions to follow use A3.)

The reason we pick a cell between rows 2 and 8 isso the formulas in the Sum and Average rows will continue to work properly. Those formulas say things like =sum(B2:B8) and =average(C2:C8). If we add new rows between what are now row 2 and row 8, the ending cells of those blocks will automatically adjust in the formulas. If we add rows at row 2 or 8, that might not happen.

Click your mouse on A3. Hold the mouse button down. Drag down until four rows are in your block.

Selects where we will be adding rows. By selecting four cells, we will add four rows. This will increase the number of rows of data from seven to eleven.

	A	B
1		X
2		100
3		200
4		300
5		400
6		500
7		600
8		700

Go up to the menu. Click on Insert, then Rows. Four rows will open up.

Check your Sum or Average rows' formulas. You should find that the blocks in the formulas have all expanded automatically. For example, the formula in B14 should now say =sum(B2:B12). It's another form of automatic adjustment that spreadsheets do. It works if inserted rows are *between* the ends of the blocks in the formulas.

If your formula is cell B14, is not =sum(B2:B12), fix it. Be sure the formulas in the rest of row 14 and 15 refer to rows 2 through 12, and *not* to row 13.

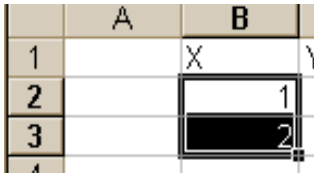
Now we can start our **analysis for the first clinic**.

Move the cell selector to B2. Type: Fills in the new data right on top of the old data.

1 Enter↵
2 Enter↵

...and so forth down to 11 in B12.

Or, for a faster way, type a 1 in cell B2 and a 2 in cell B3. Then use your mouse or ⇧Shift+↓ to highlight the block of cells B2:B3.



Fill handle ⇑

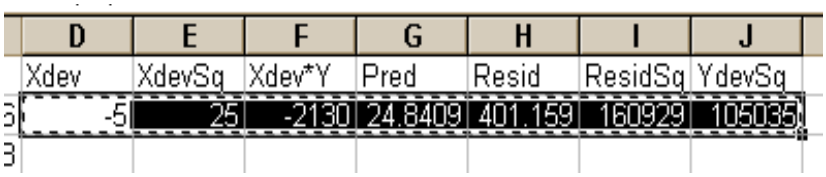
Do you see the little black square in the lower right corner of the rectangle you just created? That's the "fill handle." Click on that. Your pointer will change to a small + . Drag down to cover B12. Let go of the mouse button, and the column will be filled in. Excel figures out the pattern that you started with (1, 2) and fills in the column with 3, 4, 5, ...

Move the cell selector to C2.
Fill in column C, from C2 to C12, with the data for clinic no. 1, by typing each number and then moving down. No good short cuts here.

Some numbers in the spreadsheet will recalculate after each entry. You do not have to do anything about that.

Now, let's do columns D though J.
Move the cell selector to D2.

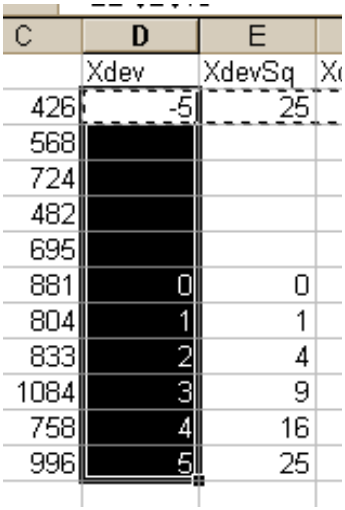
Highlight the block in the 2nd row from D2 over to J2. You can use either Shift+→, or your mouse.



	D	E	F	G	H	I	J
	Xdev	XdevSq	Xdev*Y	Pred	Resid	ResidSq	YdevSq
5	-5	25	-2130	24.8409	401.159	160929	105035
3							

Copy the selected cells to the clipboard.
(Edit Copy, Ctrl+C, or ).

Highlight the block from D2 down to D12, using Shift+↓, or your mouse.



	C	D	E
		Xdev	XdevSq
426		-5	25
568			
724			
482			
695			
881		0	0
804		1	1
833		2	4
1084		3	9
758		4	16
996		5	25

Paste from the clipboard.
(Edit Paste, Ctrl+V, or ).

The blank cells in columns D through J should fill in.

The spreadsheet recalculates. We now have the slope and intercept of the least squares line for clinic 1, as well as the test statistics s, t, and R^2.

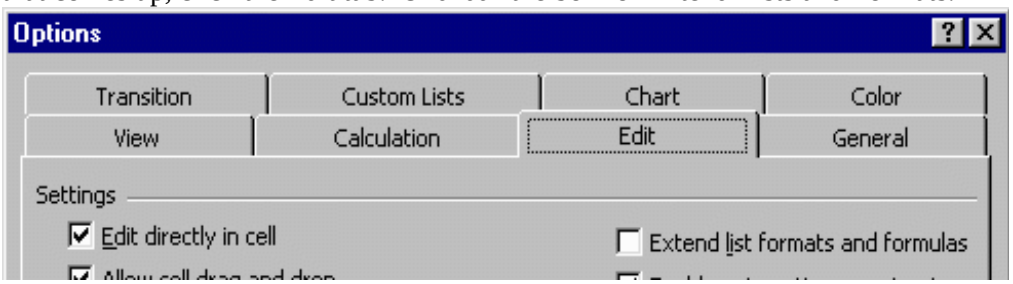
Let's add a forecast to the template.

Before we do so, **if** you are **not** in our **lab**, you may need to deal with a **dangerous feature of Excel** 2000 and 2003. (Older versions of Excel do not have this feature.) Excel 2000 and 2003's default set-up enables a feature called "Extend list formats and formulas." Because of this feature, if you now type a number in cell B13, Excel will change your formulas in cells B14 and B15 without asking you first. The ranges that those formulas sum and average will be automatically extended to B13. We do not want this. It will make our calculations incorrect.

Two ways to deal with this feature are:
1. Work around it, or
2. Disable it, so it won't bother you again. (This is what we did in the lab.)

The **workaround** is to move the cell selector to B13 and type in a letter. Any letter of the alphabet will do, as will just about any other symbol *except a number*. Press Enter after you have typed the letter. Then proceed with the instructions below this next graphic.

To **disable this dangerous feature**, select Tools from the top menu, then click Options. In the dialog box that comes up, click the Edit tab. *Uncheck* the box for Extend Lists and Formats.



Click OK to confirm the change and close this box. Then continue with the instructions that follow.

Move the cell selector to A13. Type:

`Forecast` →

A label for the forecast.

Pressing the arrow key should have taken us to B13.

B13 is where we will put the X value on which we want to base our prediction.

Type:

`12` Enter

to predict for the 12th month.

If you are using Excel 2000 or 2003 and you did not follow the precautions above, Excel will now, on its own, fill in certain cells in row 13 and change the Sum and Average formulas in B14 and B15 to make them incorrect for our purposes.

To fix this, double-click on B14, which now has the number 78, if the unwanted changes have taken place. Double-click B14 to edit the formula. Change the formula to =SUM(B2:B12) Next, double-click on B15 and change that formula to =AVERAGE(B2:B12). Changing the 13's to 12's changes the numbers in the cells back to 66 and 6.

If you have to make this correction, there is one good aspect. Excel did do one thing for us that we were going to do anyway, which is copy G column's formula down to cell G13. This puts the least squares prediction in G13. If your cell G13 is filled in this way, skip to the middle of the next page, where it says "Print the spreadsheet, using these directions:." Otherwise, continue here:

Move the cell selector to G13.

Let's put our prediction in G13, underneath the other predicted values. This will simplify including the prediction in our graphs later.

Move the cell selector to G12.
Copy this cell to the clipboard.
(Edit Copy, Ctrl+C, or ).

To avoid typing errors, we'll copy the formula from G12 to G13.

Move the cell selector to G13.

Paste from the clipboard. (Edit Paste, Ctrl+V, or ).	The forecasted value appears in G13.

Print the spreadsheet, using these directions:
(For this assignment, you can turn in the printout as described here, or you can attach your spreadsheet file to an e-mail.)

Select File Print... off of the menu or press Ctrl+P.	Either of these brings up the Print dialog box, which lets you check to see which printer Excel will use. (There are three in the lab.) You can also Preview your printout.
Click on OK to print.	The print should start. In the lab, there will be a time lag.

The print icon  starts the print immediately, with no dialog box. Use this if you are already sure the print setup is OK. (This is Excel's behavior. Other spreadsheet programs show the dialog box.)

On your printout, circle the coefficients, the prediction, Rsq, s, and t. (If you are sending in your spreadsheet file as an e-mail attachment, highlight these numbers, or add a paragraph so that I can tell that you know what these are.)

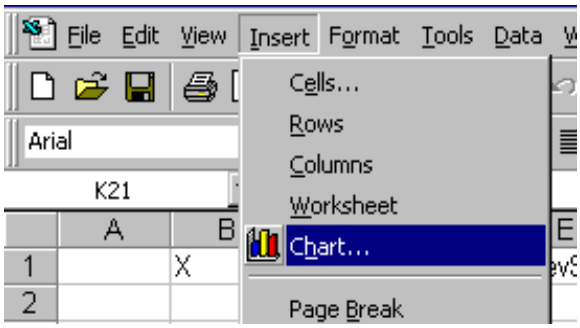
Use the t number from the spreadsheet to test the hypothesis that the true slope coefficient is 0. Describe what you are doing to make the test and what you conclude. Include a sentence that explains why you are using a two-tailed test. The mechanics of this: Compare the t number from your spreadsheet with the critical value from the t-table (in the downloadable file 716-tables for hypothesis tests.pdf). Get the critical value from the column for the α level of 0.05, and the row for at 9 degrees of freedom.

Graphs

I had you graph the data for the first assignment by hand. This time, let's get Excel to make graphs for us. (If you are using Quattro Pro and would like detailed instructions, please request them.)

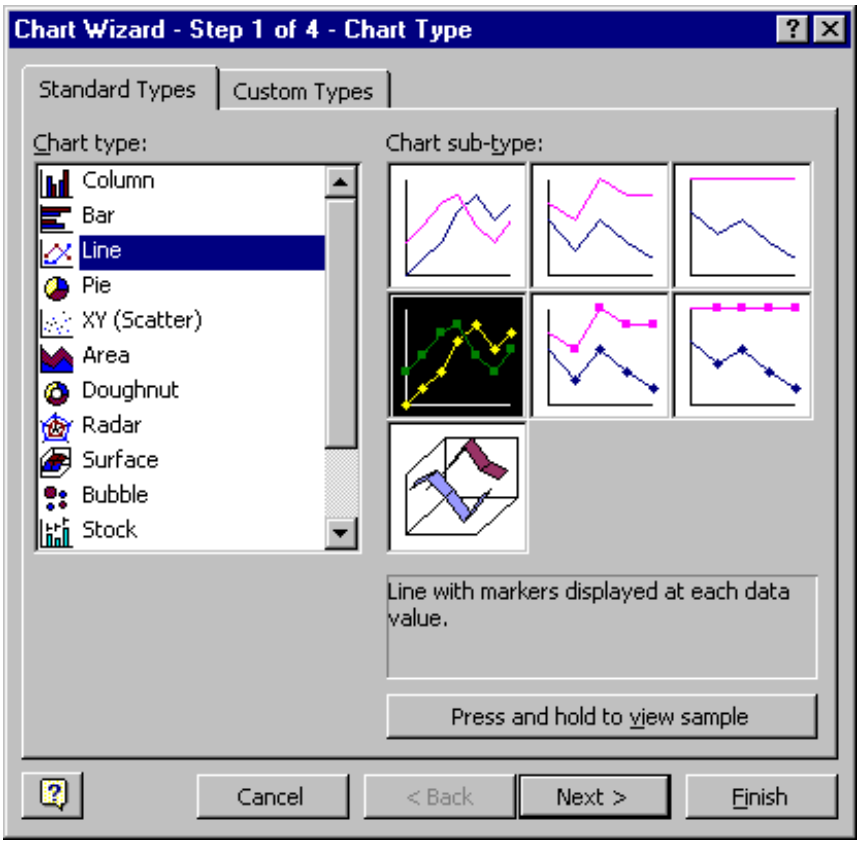
From the menu at the top of your Excel spreadsheet, select Insert, then Chart...

(Excel calls graphs "charts.")



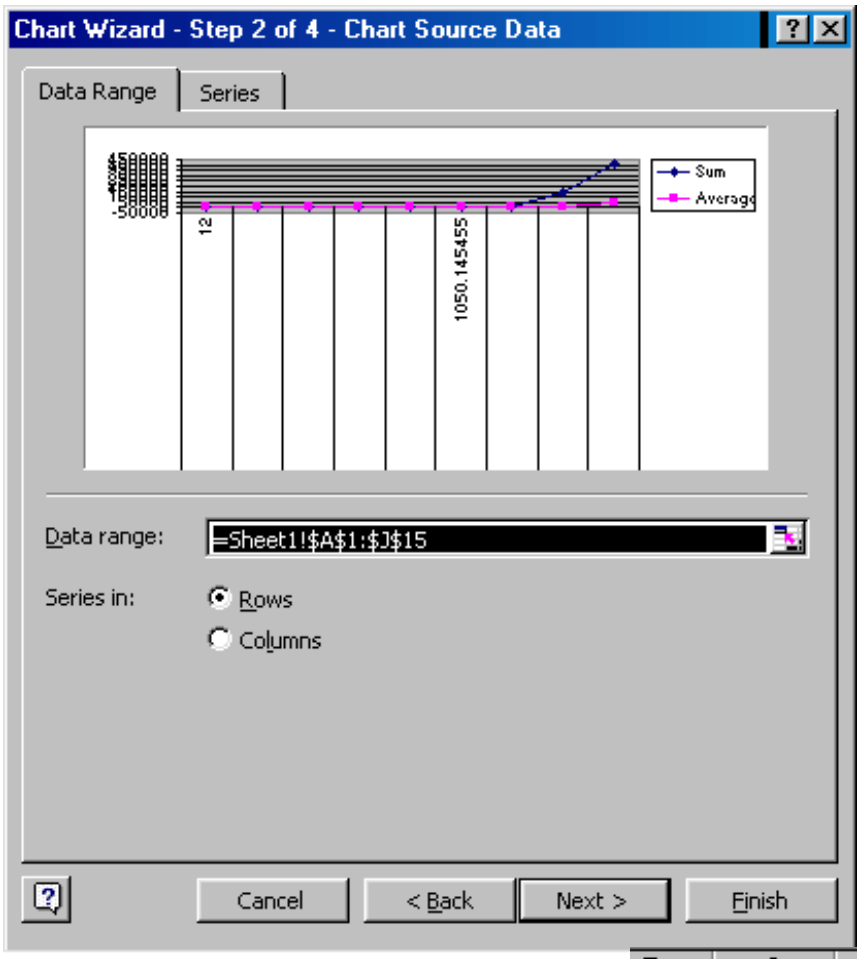
This brings up Excel's chart wizard. Select the Line Chart type, and the sub-type for "Line with markers displayed at each data value."

Then select the Next > button.



Now we tell Excel which data we want graphed. (Excel has tried to guess what data we want to graph, but it has guessed wrong. We will fix that.)

Click on the small multi-colored icon  at the right end of the Data range box. That will shrink the Chart Wizard to the top of your screen, so you can see your spreadsheet.



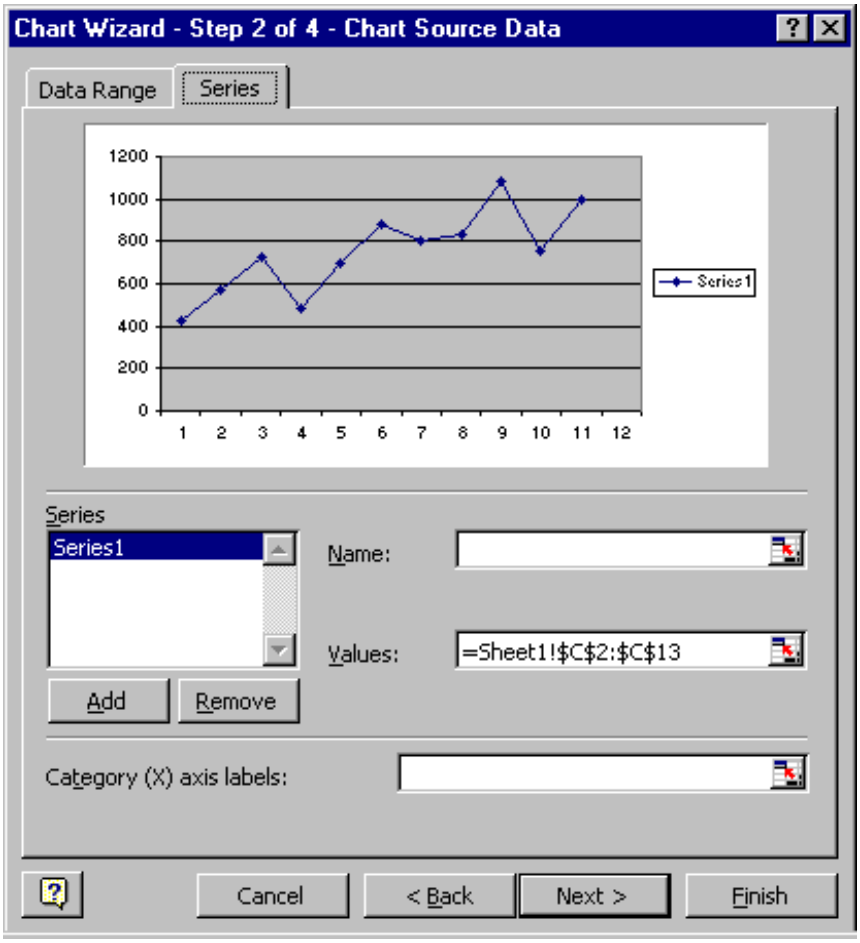
Select the block C2:C13 with the mouse or the keyboard. Notice that we include the blank cell C13. That's because we want to be able to show the predicted value for X=12 on the same graph.

	B	C	
		Y	X
1		426	
2		568	
3		724	
4		482	
5		695	
6		881	
7		804	
8		833	
9		1084	
10		758	
11		996	
12			

Press Enter ↵ when you have the block selected properly. The Chart Wizrd will come back, with a preliminary mock-up of your chart.

If your graph looks very wrong – if it has some colored dots arrayed vertically – click on the radio button for Columns.  That should fix it.

Click on the Series tab in the upper part of the window. You should see:



Click on the icon  at the right end of the box for the Category (X) labels.

Select B2:B13

B
X
1
2
3
4
5
6
7
8
9
10
11
12

and press Enter↵.

The Wizard should come back.

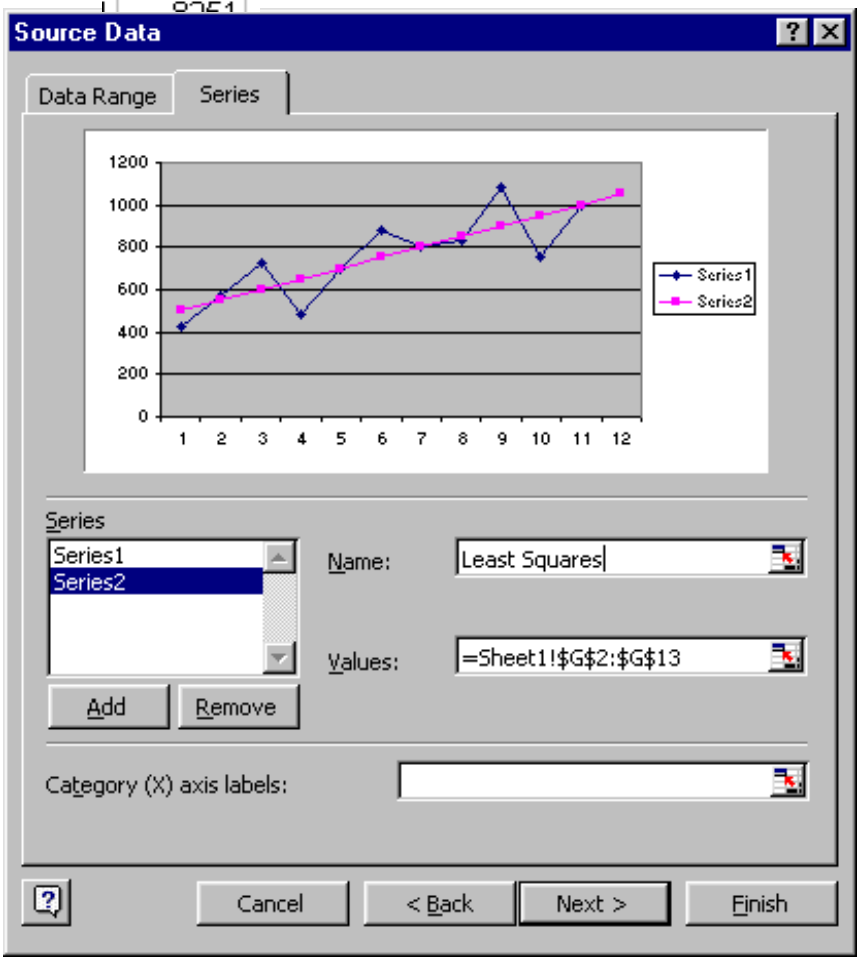
Click on the Add button on the left under Series.
Click on the icon  at the right end of the Values box.

Select the predicted values in G2:G13. Press Enter↵.

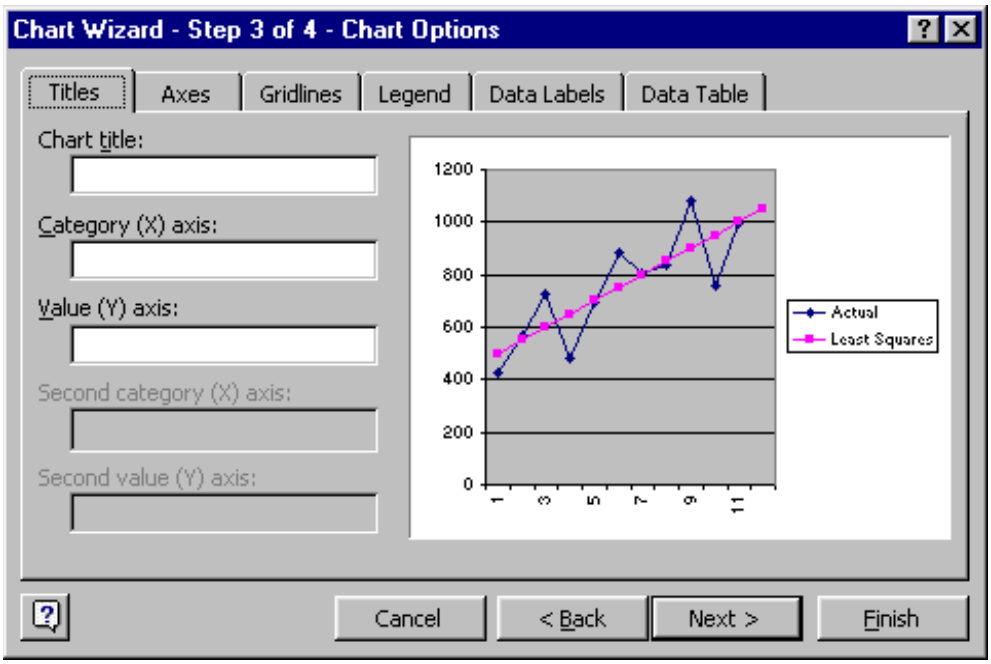
G
Pred
500.045
550.055
600.064
650.073
700.082
750.091
800.1
850.109
900.118
950.127
1000.14
1050.15

The Wizard box should come back, with a new mock-up of your graph, showing both series and the X values.

To spiff up your graph, type names for your series in the Name box. Click on Series 2 in the box on the left (if that one isn't selected already), then type "Least Squares" in the Name box.

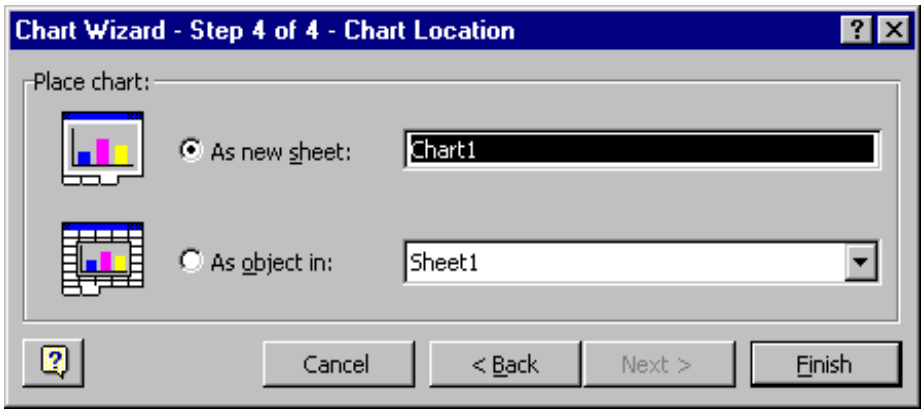


Click on Series 1 in the box on the left and type “Actual” in the Name box. The legend on your graph will change to show the names of the series.



When that’s done, click on the Next> button. That brings you to step 3.

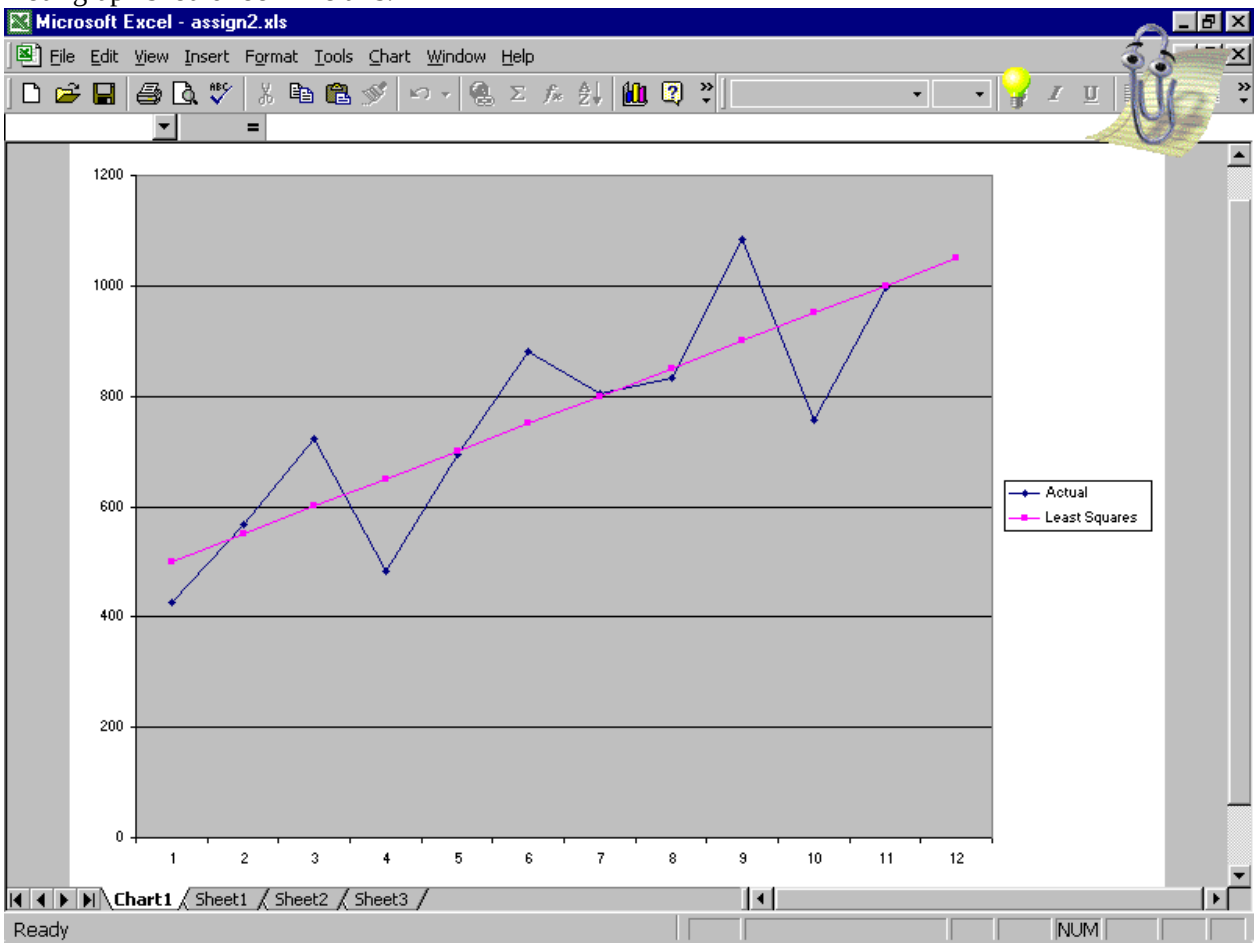
You can verify that the Legend has your series names. Other than that, here’s nothing you have to do in step 3, so click Next> to go to step 4.



For the chart location, click As new sheet.

This gives you a large chart that is on its own notebook page.

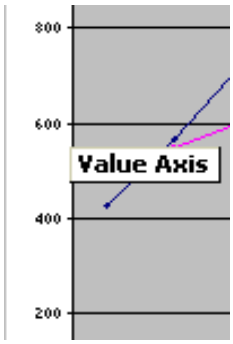
Your graph should look like this:



The tab at the bottom tells you we're on a notebook page named Chart1. Your original data are on Sheet1. Click the tabs to switch back and forth.

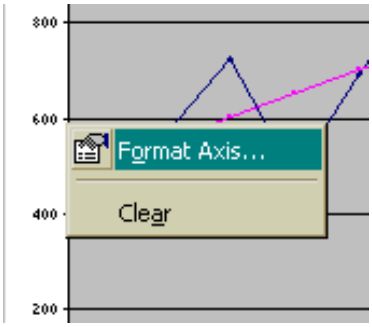
Your data points are shown as small filled diamonds, connected by line segments. The predicted values are shown as small filled squares. The predicted values form a straight line, which makes sense, since they are all on the straight line that you are fitting to the data.

Let's do something with the Y-axis. Let's make it start at 0 and go up to 1300, so it can accommodate the largest Y value we have in our data (The 3rd clinic reaches 1274 in October). This way we can compare all three clinics on graphs with the same scale.



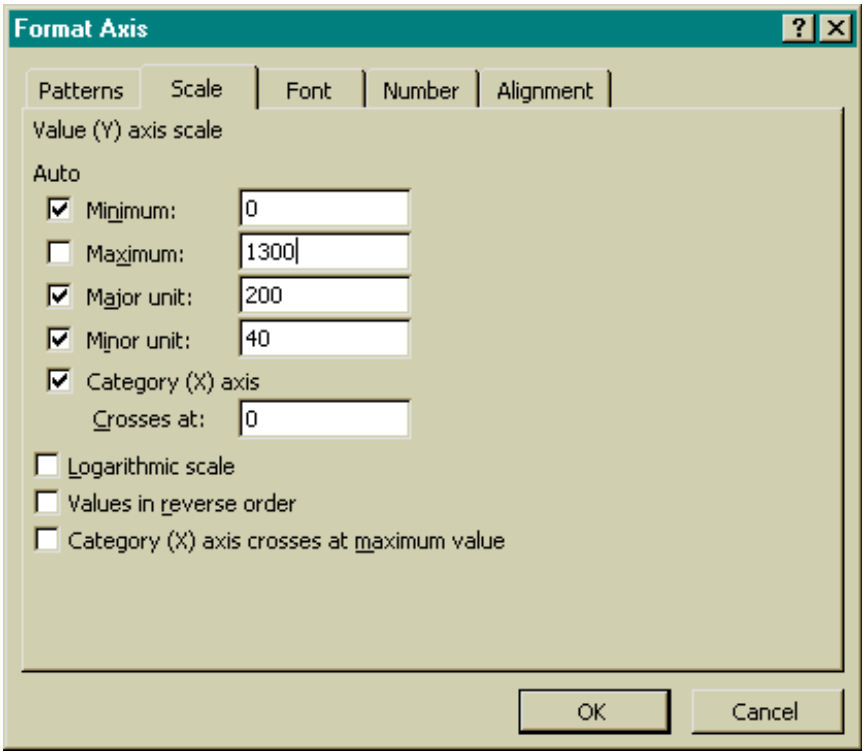
Move your mouse pointer so it is on top of the Y axis. A small sign should appear saying "Value Axis."

Right-click. A small menu should appear. Choose Format Axis... from that menu by clicking on it.

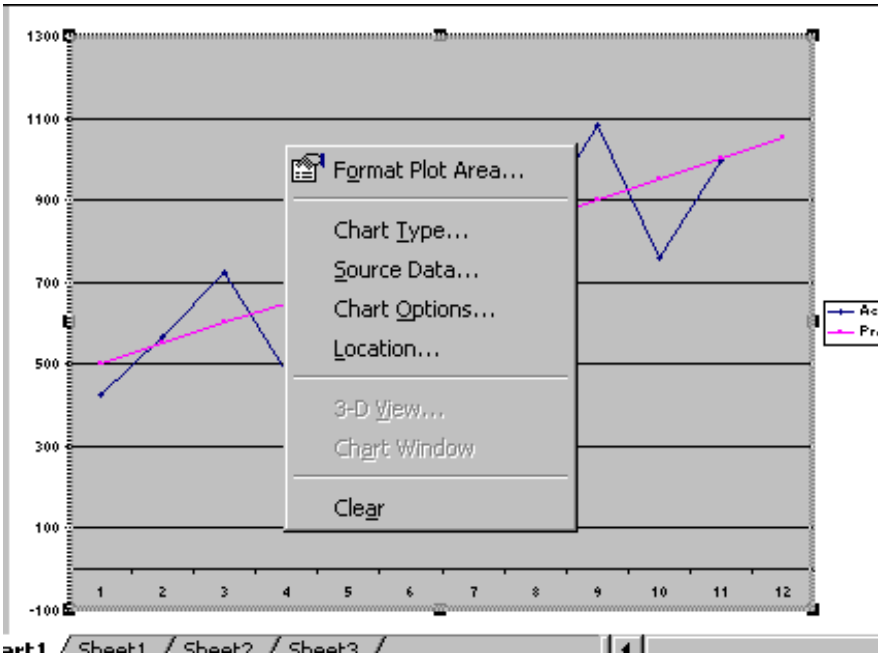


A dialog box pops up. Click on the Scale tab. Click on the box for Maximum. Type 1300 in the box for Maximum. Then click on the OK button.

The Y axis on your chart should now go up to 1300.

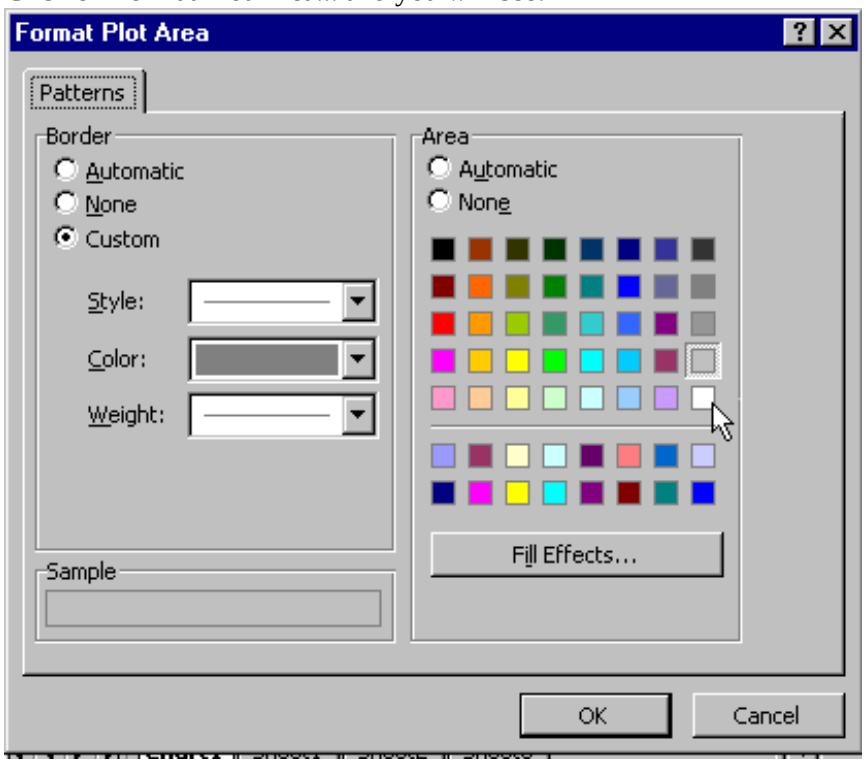


Excel allows you to customize your chart many ways. Here's one change I like to make — get rid of the grey! With the chart showing on the screen, **right**-click with your mouse in the central plot area, but not on any lines. You should see:



If you've clicked in the right place, this menu will appear, with Format Plot Area the top choice. If some other menu appears, try right-clicking in the plot area again.

Click on Format Plot Area... and you will see:



Click on the white square as shown. Then click on OK. The grey is gone!

Print your chart. To do this: Go up to the menu and select File Print. A small dialog box will appear to ask you to confirm that you want to print. Preview is there, too, if you want to verify what you are going to get before you waste paper.

This would be a good time to save your spreadsheet. To save this Assignment 2 spreadsheet separately from your Assignment 1A spreadsheet, select File, then Save As, from the menu. Type a new file name. In the lab, be sure you're saving to your floppy in A: or your storage device.

Think about what the pattern of your data points looks like and how well the least squares regression line seems to represent their overall trend. Does it seem reasonable to use linear least squares regression to make a prediction for this clinic? (Hint: The answer is Yes for this first clinic.)

Your report for clinic 1 should include:

- A printout of your spreadsheet (To get back to your spreadsheet, click on the tab for Sheet 1.)
- What the estimated slope and intercept are, and whether the slope is significantly different from 0.
- A printout of your chart
- Your prediction for Y for December (when X=12)
- Your comment about whether it seems reasonable to use linear least squares regression to make a prediction for this clinic. Judge the reasonableness by whether the assumptions at the top of the next page seem to be applicable to this clinic.

For your comments on clinics 1, 2, and 3, think about the basic assumptions needed to justify using simple linear regression:

1. The points were generated by a straight line, plus or minus a random error.
2. The error for each point has an expected value of 0.
3. The errors associated with every point tend to be about the same size. No individual points stick way out.
4. Each error is independent of all the others. In particular, each point's error is independent of the error of the point just before it.

Create graphs and reports for clinics 2 and 3.

Your reports for clinics 2 and 3 should each include:

- a printout of the spreadsheet
- what the estimated slope and intercept are, and whether the slope is significantly different from 0.
- a printout of the chart
- the least squares line's prediction for Y for December (when X=12) for that clinic
- a comment about whether it seems reasonable to use linear least squares regression to make a prediction for this clinic. Judge the reasonableness by whether the assumptions at the top of this page seem to be applicable to this clinic.

Sounds like a lot of work? It is less extra work than you might think. All you have to do is type in the clinic 2 data into column C, where the clinic 1 data are. Type right over the clinic 1 data in cells C2 through C12. The spreadsheet will recalculate and the graph will redraw with the new data. When you click on the Chart1 tab, you will see the new graph.

As you look at the graph for clinic 2, ask yourself what your eyeball prediction would be for month 12, based on the apparent pattern of the data points. Ignore the least squares line when you think about this.

Once you have made your eyeball prediction for clinic 2, compare it with the least squares prediction. Are they the same? If not, state which assumption or assumptions above appear to fail for clinic 2. This is how you justify using your eyeball prediction instead of the least squares prediction.

Now do clinic 3. Put in clinic 3's data where clinic 2's were. Again, look at the graph, which will now be different, because it will be based on clinic 3's data. Looking at the data points, and not at the least squares regression line, ask yourself what you would predict for visits in month 12.

Compare your eyeball prediction for clinic 3 with the least squares prediction. Are they the same? State which assumption or assumptions above appear to fail for clinic 3. This is how you would justify rejecting the least squares prediction in favor of your prediction, if your eyeball prediction were different from the least squares prediction.

The final piece of your assignment is this: Pick the one clinic of the three for which you think least squares is most acceptable as a prediction tool. Use the following formula to calculate the 95% confidence interval for the prediction for that clinic for December. Mark the interval on your graph for

that clinic's data.

The 95% confidence interval for the prediction is:

$$\hat{Y}_0 \pm t_{0.05}\, s \sqrt{\frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2} + 1}$$

See if you can get your spreadsheet to calculate this.
Some hints:

$\hat{Y}_0$ (read "Y-hat-sub-0") is the predicted value of Y for X_0=12. It's in the spreadsheet, at the bottom of the column of predicted values, in the row for X being 12.

$t_{0.05}$ is a number from the t-table, not from your spreadsheet. $t_{0.05}$ is the critical value for hypothesis testing at the 5% significance level in a two-tailed test. Use the row in the table for 9 degrees of freedom, because the degrees of freedom in a simple regression is N-2. (N is the number of observed points, 11.)

You multiply this by s, which you have on your spreadsheet in the summary statistics section.

You multiply next by the square root of three things added together.

1. 1/N is 1 divided by the number of data lines (11).

2. The big fraction in the middle is put together this way:

   X_0 is the value we are predicting for, which is 12.
   $\bar{X}$ is the mean of the X's, which is in the B column, in the row for averages.
   Square the difference between those to get $(X_0 - \bar{X})^2$.

   $\sum_{i=1}^{N}(X_i - \bar{X})^2$ is in the cell of your spreadsheet that is in row for sums and in the column for the X deviations squared.

3. The 1 added at the end also goes inside the square root.

Once you have the big expression after the ±, there are two numbers to calculate. These are $\hat{Y}_0$ plus the big expression and $\hat{Y}_0$ minus the big expression.

If everything worked OK, congratulations! You are well on your way to becoming a spreadsheet guru. If word gets around that you are good at spreadsheet work, you will soon have lots of friends!