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

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 will see that in this assignment.

One way to assess whether a data set is good for the least squares regression method is to draw a graph. We will use that method here. This time, we will get Excel to draw the graphs for us. We will visually inspect the graphs to see if the assumptions behind using least squares seem to 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:

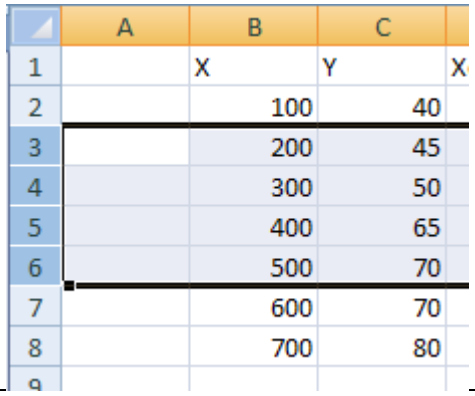
		Clinic visits		
	Month	No. 1	No. 2	No. 3
January	1	426	310	539
February	2	568	474	573
March	3	724	613	608
April	4	482	726	642
May	5	695	814	677
June	6	881	877	711
July	7	804	914	746
August	8	833	926	781
September	9	1084	913	815
October	10	758	874	1274
November	11	996	810	884

We will use our regression template from assignment 1A to calculate a least squares line for each of these clinics. First, open the spreadsheet file in which you saved Assignment 1A. To do this in the computer lab, insert your storage device. Double-click the desktop's My Computer icon, navigate to your device and to the folder containing your Assignment 1A spreadsheet. Double-click its icon. Alternatively, you can start Excel, click the Office Button and then Open, and then navigate to your file.

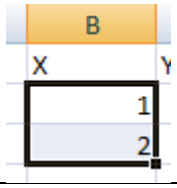
Add Blank Rows to Accommodate the New Data

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.

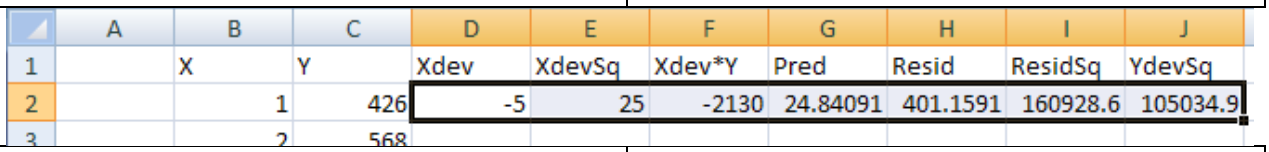
What You Do	**What It Does**
Move your mouse pointer to the 3 in the blue row numbers on the left edge of the spreadsheet. (Actually, any row from 3 through 8 will do.)	We want the formulas in the Sum and Average rows to adjust properly when we add the new rows. For example, the formula for the sum of the B column is =sum(B2:B8). If we add new rows at row 2 or after row 8, the range for the sum will not adjust to include the new rows.

Click your mouse on 3. Hold the mouse button down. Drag down until four rows (3 through 6) are selected. 	Selects where we will be adding four rows. This will increase the number of rows of data from seven to eleven.
Right-click. Select Insert from the pop-up menu.	Four rows will open up.

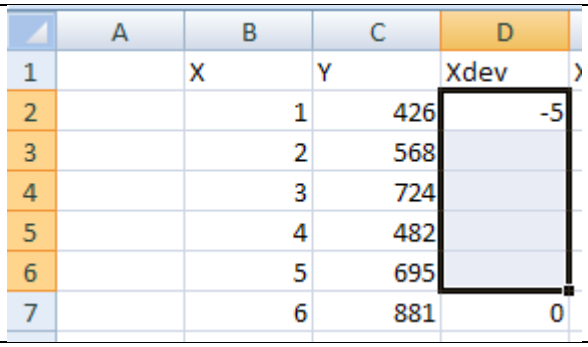
Check your Sum or Average rows' formulas by clicking on one or two of those cells or using Ctrl+` (again, that is the tilde, which is to the left of the 1 key on many keyboards) to show all the formulas. You should see that the blocks in the formulas have all expanded automatically. For example, the formula in B14 should now say =sum(B2:B12). It's another form of automatic adjustment that spreadsheets do. It works if the inserted rows are *between* the ends of the blocks in the formulas.	
Now we can start our **analysis for the first clinic**.	

Move the cell selector to B2. Type: 1 Enter 2 Enter ...and so forth down to `11` in B12.	Fills in the new data right on top of the old data.
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.	 Do you see the little black square in the lower right corner of the rectangle you just created? That is the "fill handle." ←
Click on the fill handle. Your pointer will change to a small + . Drag down to cover cell 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 by pressing Enter. 426 568 724 482 695 881 804 833 1084 758 996	Some numbers in the spreadsheet will recalculate after each entry. You do not have to do anything about that.
Move the cell selector to D2.	Let's fill in the blank spaces in columns D though J.
Highlight the block in the 2nd row from D2 over to J2. You can use either Shift+→, or your mouse.	



	A	B	C	D	E	F	G	H	I	J
1		X	Y	Xdev	XdevSq	Xdev*Y	Pred	Resid	ResidSq	YdevSq
2		1	426	-5	25	-2130	24.84091	401.1591	160928.6	105034.9
3		2	568							

Copy the selected cells to the clipboard. (Ctrl+C or click the Copy icon under the Home tab on the menu).	
Select the block from D2 down to D6, using Shift+↓, or your mouse. (If you inserted your new rows further down, extend the selected block in the D column so that it covers all of the empty cells in the D column.)	
Paste from the clipboard. (Ctrl+V or the Paste icon under the Home tab.	The blank cells in columns D through J should fill in.

	A	B	C	D
1		X	Y	Xdev
2		1	426	-5
3		2	568	
4		3	724	
5		4	482	
6		5	695	
7		6	881	0

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.

Now to add a **forecast**.

Before we do so, you may need to deal with a **dangerous feature of Excel**, introduced with Excel 2000. (I consider it dangerous, anyway). Excel's default set-up enables a feature called "Extend data range formats and formulas." (It's "Extend list formats and formulas" in older versions.) 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. (It's OK to disable this on a HSPM lab computer. We have done this in the past, but the fix may have been undone when the newer Excel version was installed. Outside of the lab, if you are using a computer that you do not own, do the workaround instead.)

The **workaround** is to move the cell selector to B13 and type in a letter. Any letter of the alphabet will do. Just *don't type a number*. Press Enter↵ after you have typed the letter. Then proceed with the instructions below the next paragraph.

To **disable this feature**, click the Office Button. Click the small rectangular Excel Options button. Click Advanced. *Uncheck* the box for Extend data range formats and formulas. Click OK.

> In older Excel versions, disable the feature by selecting Tools from the top menu, then clicking Options. In the dialog box that comes up, click the Edit tab. *Uncheck* the box for Extend Lists and Formats. Click OK.

Move the cell selector to A13. Type: `Forecast` →	A label for the forecast.
You should now be in B13.	B13 is where we will put the X value for which we want the predicted Y value.
If you followed the workaround instructions above, there is a letter in cell B13. You now type over that letter. Type: `12` Enter↵	to predict visits for the 12th month.

> If you did not do the workaround, and if the feature I don't like has not been disabled, 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. B14 will now say 78, for example. ☹
>
> Don't worry! Ctrl+Z will fix things. ☺ Then do the workaround: Type a letter in cell B13. Press Enter↵ . Move back to B13 and type 12 Enter↵ .

Let's put our predicted visits for month twelve 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. (Ctrl+C or the Copy icon under the Home tab).	We'll copy the prediction formula from G12 to G13.
Move the cell selector to G13. Paste from the clipboard. (Ctrl+V or the Paste icon under the Home tab).	 The forecasted visits for month twelve appears in G13.

In your write-up, tell me what your prediction is for clinic 1. That's the number in G13. Also report your Rsq, s, and t. (Your write-up can be a word processor file, or you can add comments to your spreadsheet if you know how to do that.)

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 http://sambaker.com/courses/J716/pdf/716-tables for hypothesis tests. Get the critical value from the column for the α level of 0.05, and the row for at 9 degrees of freedom.

Type the t-value from the t-table in cell D18, right under the calculated t-value in C18. That way, you'll have both numbers in front of you. That will also set you up for the calculation of the confidence interval of the slope coefficient. It might be a good idea to type `95% t from table` in cell D19.

I recommend that you verify that your work is correct so far, using the online answer checker at http://sambaker.com/courses/J716/a02/a2.html

Slope coefficient confidence interval

The slope coefficient confidence interval tells you which possible values for the slope you would accept and which you would reject, based on the t-test.

In cell E16, type `Confidence interval`. (By now, you know to press Enter↵ to finish a cell entry.)

Just below, in cell E17, type `=D18*C17+B17`
This gives you the top end of the slope coefficient's 95% confidence interval. Notice that the formula uses the number from the t-table in cell D18, not the calculated t number in D17.

16			Std Err	t coeff=0	confidence interval
17	Slope	50	11.796	4.2386	=D18*C17+B17
18	Intercept	450.09		2.262	
19				95% t from table	

In cell F17, type `-D18*C17+B17` This formula starts with a minus sign, not an equal sign.
This gives you the low end of the 95% confidence interval for the coefficient. (It is 95% because the number in cell D18 is from the 95% or .05 column of the t-table.)

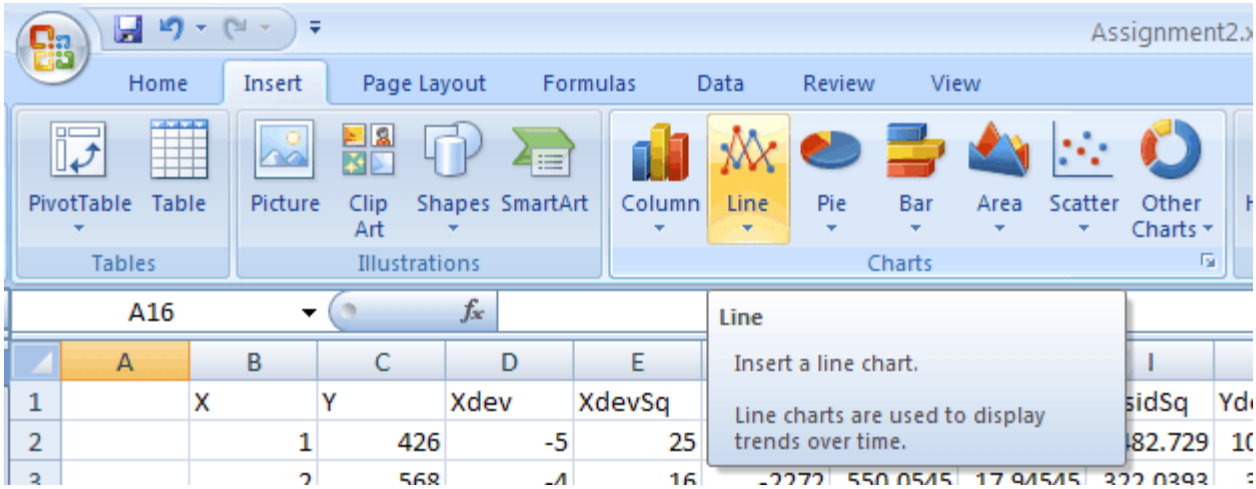
Zero is not in the confidence interval, right? If the calculated t number, in D17, is greater than the t-table number, in D18, then zero will not be in the confidence interval for the slope.

With 95% confidence, do you think that the true slope might be 80? How about 40?

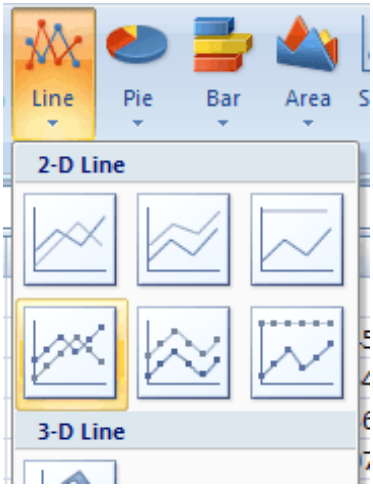
Graphs

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 need instructions for older versions of Excel, please ask.)

From the menu at the top of your Excel spreadsheet, click the Insert tab. In the Chart section, choose Line chart.

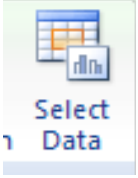


In the menu that pops up, pick Line with markers.

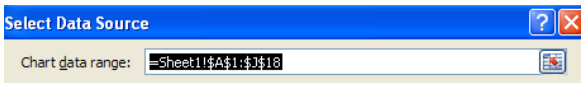


A chart will appear that looks funny. We need to tell it which data to display, because it is evidently not good at guessing our intentions.

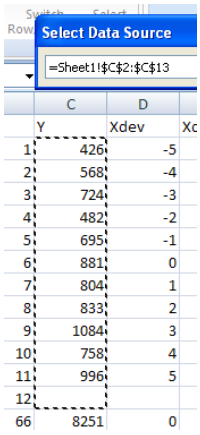
Click on



Click on  in the Select Data Source box at the right end of the Chart data range line.



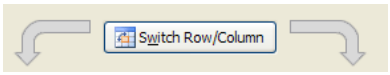
That will shrink the Select Data Source box, so that you can see your spreadsheet.

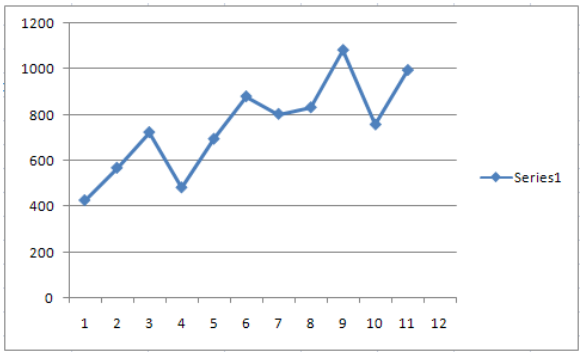


Select the block C2:C13 with the mouse or the keyboard. Notice that we include the blank cell C13. This will make room on the graph for the predicted value for X=12.

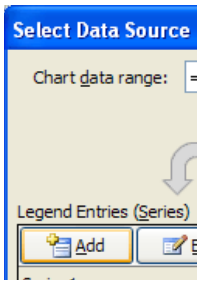
Press Enter when you have the block selected properly. A new mock-up of your chart will show.

If your graph looks very wrong – if it has colored dots arrayed vertically – click on the button in the Select Data Source box that switches rows and columns. The graph should then change to something like this:

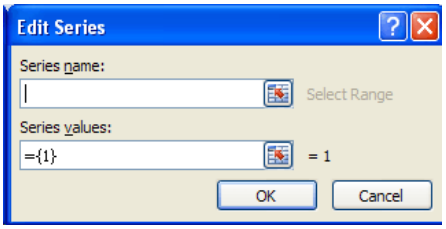




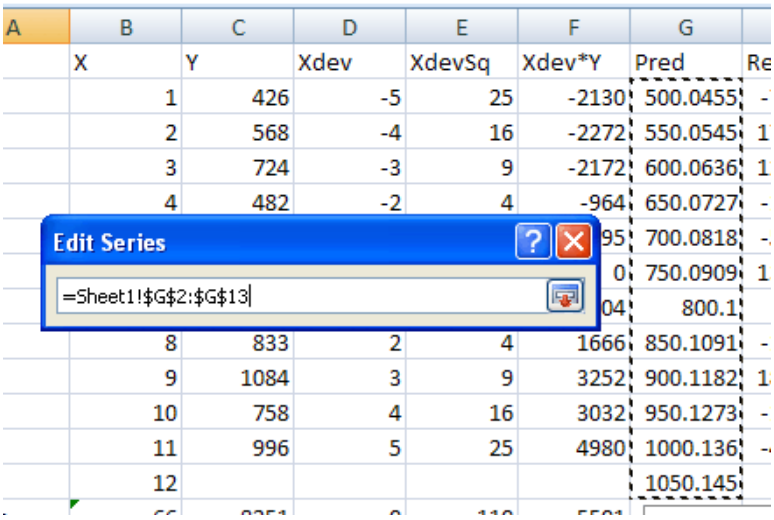
We'll add the predicted values to this graph. Click the Add button in the Select Data Source box. The button's border turns yellow when your mouse pointer is on it.



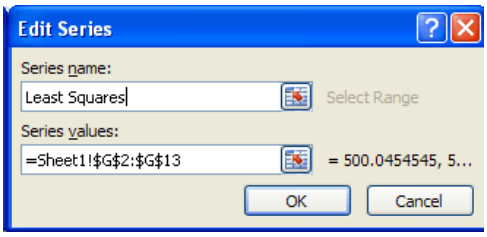
A box like this will pop up:



Click on the lower symbol-with-the-red-arrow. Move the dialog box away from column G. Then click on G2 and drag with your mouse down to G13.



Press Enter and the mock-up chart should change to show both the actual Y values and the predicted Y values. The predicted values should form a straight line.



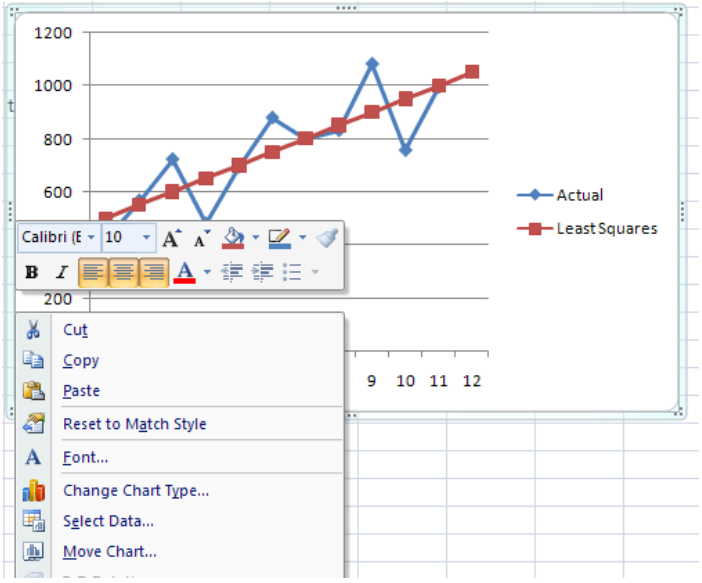
While the Edit Series box is still on the screen, click in the Screen name box and type `Least Squares`

This will label the predicted value line "least squares." This is to remind you that the prediction line is drawn using the least squares method.

Click OK to effect the change.

Click on Series 1 and then the Edit button. Type `Actual` in the Series name box. Click OK.

You now have a chart! Unfortunately, it's in the way. You can click near the chart's edge to move it. I recommend that you put the chart on its own spreadsheet tab. To do this, right-click just inside one edge of the chart. If you find the right place to click, the border of the chart will turn blue and a pop-up menu will appear.



Click on Move Chart...

Click the radio button for New Sheet and click OK.

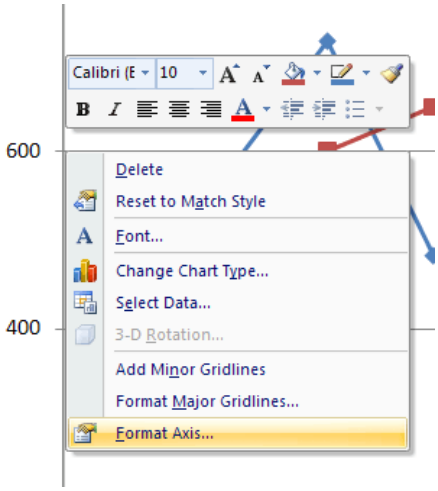
This gives you a large chart that is on its own page. Tabs at the bottom of Excel's window let you switch between seeing the chart ("Chart1") or the data ("Sheet1").

Your data points are shown as small filled diamonds, connected by line segments. The predicted values are shown as small filled squares. As mentioned, the predicted values form a straight line, because that is what the least squares method does – it draws a straight line.

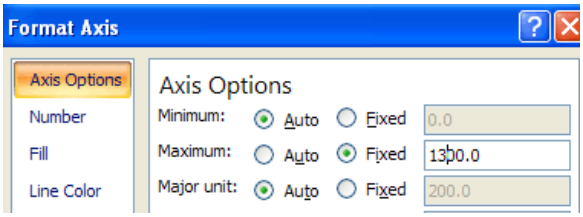
Let's change the Y-axis. Let's make it start at 0 and go up to 1300, so it can accommodate the largest Y value we have in our data (The 3rd clinic reaches 1274 in October). This way we can compare all three clinics on graphs with the same scale.

On Chart1, right-click on the Y-axis. If you find right place, the pop-up menu shown to the right will appear.

Click on Format Axis...



This brings up the Format Axis dialog box.



Click on Fixed and type 1300 in the box. Then click Close.

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

Take a moment to admire your chart!

This would be a good time to save your spreadsheet. Save this Assignment 2 spreadsheet separately from your Assignment 1A spreadsheet. Click the Office Button and select Save As. In the lab, be sure you're saving to 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.)

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 have the same variance. The distances from the points to the line are all in a size range that is even across the data set. 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.

Your report for clinic 1 should include:

- What the estimated slope and intercept are, and whether the slope is significantly different from 0.
- 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 in the box on page 10 seem to be applicable to this clinic.

Print the graph for clinic 1, so you can look at it later. To do this: Click on the Chart1 tab, so that the graph is showing on the screen. Click on the Office Button, then Print, then Print. Under Print What, print the Active sheet.

Clinics 2

Now for clinic 2! You'll be typing clinic 2's data in column C, overwriting clinic 1's data. In case you want to put clinic 1's data back, first copy clinic 1's data to Sheet 2.

Select cells C2 through C12 with your mouse or ⇧Shift and an arrow key.

Use Ctrl+C or the Copy icon under the Home tab to copy those cells to the clipboard.

Click on the Sheet2 tab at the bottom of the screen.

Click on cell A2. (That's just my suggestion. You can use any cell.)

Use Ctrl+V or the Paste icon under the Home tab to paste from the clipboard. Clinic 1's data should appear.

Click on the Sheet1 tab to go back to the sheet with your regression template.

Type clinic 2's data into column C, where clinic 1's data are.
Type right over the clinic 1 data in cells C2 through C12. The spreadsheet will recalculate as you go. If you have gotten good and copying and pasting, copy the data below and paste into cell C2.

```
310
474
613
726
814
877
914
926
913
874
810
```

Report in your write-up:

- what the estimated slope and intercept are, and whether the slope is significantly different from 0.
- the least squares line's prediction for Y for December (when X=12) for that clinic
- How do the Rsq, s, and t of this clinic compare with the Rsq, s, and t of clinic 1? Based on those statistics, does the least squares line fit Clinic 2's data as well as it fits clinic 1's data?

Click on the Chart1 tab. The graph will now be showing clinic 2's data.

- If the least squares line were not on the graph, what would your intuitive prediction be for the next month?
- Now, mentally put the least squares line back on your graph. How does your intuitive prediction compare with the prediction from the least squares line?
- If your intuitive prediction differs from the least squares prediction, justify your intuitive prediction telling me which assumption in the box on page 10 seems implausible for clinic 2. (The point is: If you suspect that least squares is not the best prediction method, then one of those assumptions must be wrong. If none of those assumptions are wrong, then least squares is the best prediction method.)

You may wish to print the graph for clinic 2, so you will have it after you put in clinic 3's data.

Clinic 3

Now for clinic 3. First, copy clinic 2's data Sheet2, next to clinic 1's data. (This is optional, but a good idea if you want to go back.)

Type in (or copy and paste) clinic 3's data where clinic 2's were. Here are clinic 3's data:

```
539
573
608
642
677
711
746
781
815
1274
884
```

Report in your write-up:

- what the estimated slope and intercept are, and whether the slope is significantly different from 0.
- the least squares line's prediction for Y for December (when X=12) for clinic 3.
- How do the Rsq, s, and t of this clinic compare with the Rsq, s, and t of clinic 1? Based on those statistics, does the least squares line fit Clinic 3's data as well as it fits clinic 1's data?

Click on the Chart1 tab. The graph will now be showing clinic 3's data.

- If the least squares line were not on the graph, what would your intuitive prediction be for the next month?
- Now, mentally put the least squares line back on your graph. How does your intuitive prediction compare with the prediction from the least squares line?
- If your intuitive prediction differs from the least squares prediction, justify your intuitive prediction telling me which assumption in the box on page 10 seems implausible for clinic 3.

Prediction confidence interval

The final piece of this assignment is: 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.

The 95% confidence interval for the prediction is:

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

See if you can use your spreadsheet to calculate this. Some explanation:

$\hat{Y}$ (called "Y-hat") is the predicted value of Y for X_p=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 the number from the t-table. $t_{0.05}$ is the critical value for hypothesis testing at the 5% significance level in a two-tailed test. This number is already in your spreadsheet.

You multiply this by s, which is also already in your spreadsheet.

Put a formula for the big square root mess in a cell of its own. It is the square root of the sum of three things. You might want to put each of those things in separate cells:

1. The first item is just a 1.

2. 1/N is 1 divided by the number of data lines (11).

3. The big fraction can be put together this way:

   X_p is the value we are predicting for, which is 12.

   $\overline{X}$ 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 $(X_p - \overline{X})^2$, the numerator of the fraction.

> The denominator of that big fraction is in the cell of your spreadsheet that shows the sum of the X deviations squared. It's in the Sum row under the column for XdevSq.

Take the square root of the sum 1, 1/N, and the big fraction. Then, use another cell to multiply by $t_{0.05}$ and s. That cell will then have the big expression after the ±.

In two more cells, calculate

$\hat{Y}_0$ plus the big expression and $\hat{Y}_0$ minus the big expression. ($\hat{Y}_0$ is in your spreadsheet, in row 13 of the Pred column.)

That's your 95% confidence interval for the prediction!

If you would like more guidance, see the online Answer Checker.

If you can, mark the interval on your graph for that clinic. The marks will be above and below the right end of your regression line. If you can't do this on the screen, please print a copy of one of the graphs and mark on the paper. (You don't have to scan the paper for submission.) I would like you to see how wide the 95% prediction confidence interval is.

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 spreadsheets, you will have lots of friends!

Submit on Blackboard your spreadsheet file and your write-up.