Assignment 1A. Predicted Values, Residuals, and Measures of Goodness of Fit[1]
© 2011 Samuel L. Baker

This script shows how to extend the spreadsheet from Assignment 1 so it can:

1. Calculate a predicted value of Y for an X value that you choose.
2. Calculate the residuals for each observation.
3. Use the residuals to calculate measures of how well the regression line fits the observed points.
4. Test the hypothesis that the true slope is 0.

You will be turning in a copy of this extended spreadsheet.

Retrieving your spreadsheet

Start Excel and open the file in which you saved for Assignment 1. Do this with the round button and then Open or by pressing Ctrl+O. In the dialog box that comes up, navigate to the folder or device that has your file. Select it.

Your saved spreadsheet from Assignment 1 should now be on your screen.

Here is what the spreadsheet will look like when we get done. (Many of your numbers will differ, as you know, because your data are different from mine.)

	A	B	C	D	E	F	G	H	I	J
1		X	Y	Xdev	XdevSq	Xdev*Y	Pred	Resid	ResidSq	YdevSq
2		100	40	-300	90000	-12000	38.10714	1.892857	3.582908	361
3		200	47	-200	40000	-9400	45.07143	1.928571	3.719388	144
4		300	43	-100	10000	-4300	52.03571	-9.03571	81.64413	256
5		400	67	0	0	0	59	8	64	64
6		500	62	100	10000	6200	65.96429	-3.96429	15.71556	9
7		600	72	200	40000	14400	72.92857	-0.92857	0.862245	169
8		700	82	300	90000	24600	79.89286	2.107143	4.440051	529
9										
10	Sum	2800	413	0	280000	19500	413	1.42E-14	173.9643	1532
11	Average	400	59	0	40000	2785.714	59	2.03E-15	24.85204	218.8571
12		Coefficier	Std. Err.	t coeff=0				Regression statistics		
13	Slope	0.069643	0.011147	6.247562				s	5.898547	
14	Intercept	31.14286						R-squared	0.886446	

---

[1] Proper grammar might be Wellness of Fit, but that sounds too much like an exercise program.

Step-by-step to to produce this spreadsheet:

What you type	**What it does**
Move the cell selector to the top of column G, cell G1, and type: `Pred` ⮕	"Pred" will be the column heading for the predicted values of Y. The predicted values come from plugging the observed X values into the regression formula.

Some textbooks call the predicted values the "fitted" values.

The predicted values come from applying this formula

Predicted Y = Intercept + Slope*X

to each of the X values. To reiterate, the predicted values here are the least squares predicted values, because they are calculated using the least squares slope and intercept.

Let's do the rest of the column headings now:

The selected cell should now be H1. In that cell, type `Resid` ⮕	Resid is for the residuals.

The residuals will be calculated by applying this formula,

Residual = Actual Y - Predicted Y

to each X value.

In I1, type: `ResidSq` ⮕	The squares of the residuals will go in this column. We need to add these up as part of the goodness of fit formula.
In J1. type: `YdevSq` ⬅	The squares of the Y deviations (the Y deviations are each Y value minus the mean of the Y's) will go in this column. The sum of these also goes into the goodness of fit calculation.

The R^2 formula uses the sums of columns I and J to measure of how well the line fits the points.

Let's fill in column G. Go to G2.

Type:

`=b14+b13*b2` ‎`Enter ↵`

The predicted value of Y is Intercept+Slope*X. B14 is the intercept. B13 is the slope. B2 is the first X value.

Spreadsheets follow the standard math convention of multiplying before adding. This formula multiplies what's in b2 by what's in b13, then adds the product to what's in b14.
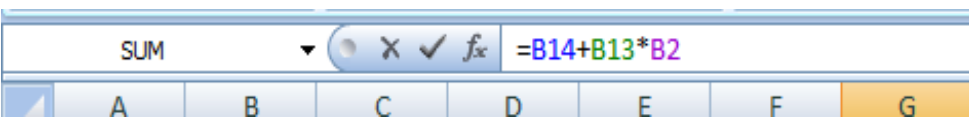
We intend to copy G2 to the rest of column G. But first we have to put in some $ signs.

Why $ signs? As we copy this cell's formula down the column, we want only the b2 to change, to b3, b4, etc. We do not want the b13 and b14 to change. We want them to stay right on the cells for the slope and intercept. Putting $ signs in the B13 and B14 is how we tell Excel to not adjust those cell references when we copy the formula to other cells.

Here is how to put in the $ signs:

Move the cell selector to G2 if it's not still there.
Press `F2` or click on the formula in the box above the column headings to edit the entry in the current cell, G2.

SUM	▾	X ✓ ƒx	=B14+B13*B2			
A	B	C	D	E	F	G

Click with your mouse or press `←` a few times to move the cursor on top of the `B13`. On a Windows computer, you can put the cursor on the B, 1, or 3. It doesn't matter which. When it's there, press:

We are editing the cell to add the $ signs to the formula where they are needed so that it will copy correctly.

`F4` to make the B13 turn into $B$13.

This way, the slope reference, B13, will not change in copies of the formula.

On a Mac, command+T should do it.

Give `B14` the same treatment, by moving the cursor there and pressing `F4` to change it to $B$14 or typing in the $ signs.

This makes the intercept reference stay still, too.

The formula should now look like this:
`=$B$14+$B$13*B2`
If you're in the cell, you can press `Enter ↵`

This completes the editing of the formula.

Copy the formula down the column from G2
through G8. Here's how:

Move the cell selector to G2. Copy that cell to
the clipboard by clicking Copy or by pressing
Ctrl +C.

We want to copy the formula in G2.

Select the destination block of cells G2:G8 with
⇧ Shift and ↓ or by left-clicking and dragging
with the mouse.

F	G	
ev*Y	Pred	F
-12000	38.10714	
-9400		
-4300		
0		
6200		
14400		
24600		

Paste from the clipboard by clicking Paste or
with Ctrl +V.

The box in the picture above is filled in with
predicted values.

The residuals are next. Move the cell selector to
H2. Type an equals sign:

Each residual is the actual Y minus the predicted
Y. There is one residual for each row of data in
the table.

=

Once that = is typed, Excel is expecting a formula. We could just type the formula, as we've been doing,
but there's another way that's more reliable. Excel allows us to use the cell selector to pick out cell
references for formulas.

Click with your mouse on C2, or press
← ← ← ← ←
to move the cell selector to C2.

The cell entry back in H2 should say =C2.

Now press the minus sign (or the hyphen):

–

 Our formula-in-process becomes =C2– and the
cell selector jumps back to H2.

Click with your mouse on G2, or press ←
to pick up G2.

The formula in H2 becomes =C2-G2. This is
the actual Y (from C2) minus the least-squares
predicted Y (from G2).

Press:

Enter ↵

and the first residual appears in H2.

Let's fill in I2, the square of that residual, using the same method of pointing to the cells we want in our formula.

Click on I2, or press → to move the current cell to I2.

Press:
=

The equals sign starts a formula. Now we can point to the cells we want in the formula.

Click on H2, or press ←

Puts H2 in the formula.

Press
*

Puts * in the formula, and moves the cell selector back to I2.

Click on H2, or press ←

Cell I2 now has =H2*H2 in I2.

Press:

Enter ↵

and the square of the residual appears in I2.

Now let's do J2. Move the cell selector there with:

→

Column J is for the squares of the Y deviations. We'll skip the intermediate step of filling in a separate Ydev column.

Type an equals sign and an open parenthesis:

= (

Starts the formula.

Click on C2, or press ← seven times to pick up C2.

Uses the mouse or the arrow keys to "point" to cell C2 for inclusion in the formula.

Press:
−

to put in a minus sign.
The cell selector moves back to J2 by itself.

Use the arrows or the mouse to move the cell pointer to C11.

The average of the Y's is C11.
The formula, so far, is now =(C2-C11

Press:
F4   or type in $ signs to change C11 to
        $C$11.

This designates C11 as a cell reference that should not change when we copy the formula in J2 to another cell.

Now close the parentheses:

)

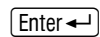The formula so far calculates the first Y value minus the Y mean.

To get the square of this, type:

^2

Your formula should now be:
=(C2-$C$11)^2

(Quattro Pro may add an $A: to this formula, indicating the first spreadsheet page.)

When you have it right, press:

Enter ↵

This completes the cell entry.

You now have entries in H2, I2, and J2.

We can fill in the H, I, and J columns all at once by copying the block H2:J2 down to rows 3 through 8.

Move the cell selector to H2. Then ...

Hold down ⇧ Shift and use the arrows or the mouse to move to J2. The selected block should be H2:J2.

Selecting a block across row 2 that you can use to fill in the rows below.

Copy the cells to the clipboard by clicking Copy or by pressing Ctrl+C.

Select the block of cells H2:H8 with ⇧ Shift and ↓ or by left-clicking and dragging with the mouse.

Paste from the clipboard by clicking Paste or by pressing Ctrl+V.

All the blank spaces in the top part of the table should fill in.

Next, we need to copy our sum and average formulas to the empty spaces in new data columns.

Move the cell selector down to F10. Then press:

⇧ Shift+↓   extends the selected block down to F11.

Getting ready to copy and paste the column sum and average formulas.

A note on how I'm using the + sign in these instructions: Windows documentation uses the + sign with shift keys to indicate that one key is held down while the other is pressed. For example, ⇧ Shift+↓ means hold down the shift key while pressing the arrow key. It doesn't mean to press +. The trouble is, my instructions also use the + sign to indicate places where you actually want a plus sign, such as in formulas like =B14+B13*B2. Hopefully, the context will make it clear what a "+" means.

Copy this block to the clipboard by clicking Copy or by pressing Ctrl+C.

The sum and average formulas are now in the clipboard.

Click again on F10.

⇧Shift+→ five times to make F10:J10 the selected block.

We are copying a vertical block across a row, so we designate just the top row of the area we want filled in.

Paste from the clipboard, by clicking Paste or with Ctrl+V.

The block F10:J11 should fill in with sums and averages for each column.

Let's take a minute to look at the numbers that Excel just calculated.

The average of the predicted values is the same as the average of the Y values. This is because the regression line always goes through the mean, or center of gravity, of the data points.

The residuals sum to 0, or some number like 1.42E-14 that is very close to 0. This reflects one of the assumptions on which the derivation of the regression line formula is based, namely that the expected value of each error from the true line is 0. (As explained in Assignment 1, a number like 1.42E-14 means -0.0000000000000142. If the sum of your residuals is this small, then what you are seeing is just cumulative round-off error. Your formulas are correct. If the sum of your residuals isn't this minuscule, check your formulas.)

Three goodness-of-fit statistics: R^2, s, and t

With the sum of the squares of the residuals and the sum of the squares of the Y deviations, we can figure out the R-squared, the most popular measure of goodness of fit. We can also figure out the standard error ("s"), and the Student's t statistic for testing the hypothesis that there is no linear relationship between Y and X.

Go down to H12 and type
Regression Statistics

This labels what will be below here as statistics that apply to the whole regression equation.

Go down to H13 and type s.
In H14, type R-squared

Puts in labels for the formulas we are about to create.

Move the cell selector to I14.

This is where we'll put the R-squared formula. Our spreadsheet is set up so that this is just 1-I10/J10.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Start by typing:

The first part of the R^2 formula.

`=1-`

Use the arrow keys or the mouse to move the cell selector to I10.

In the fraction, the numerator is the sum of the squares of the residuals. This is in cell I10.

Then press:

`/`

/ is the symbol for division

Click on cell J10, or use the arrow keys to move the cell selector to J10.

The denominator of the fraction is the sum of the squares of the Y deviations from the Y mean, which is in J10.

The formula should read `=1-I10/J10` .
When it does, press Enter↵.

The R-squared formula is complete!

The R-squared tells us how the spread of the Y values around the regression compares with the spread of the Y values from their average.

If the Y values are a lot closer to the regression line than they are to their average, the line fits good (well). In that case, I10/J10 will be close to 0, so the R-squared (1-I10/J10) will be near 1.

If the Y values are just about as far from the line as they are from their average, I10 will be almost as big as J10. I10/J10 will be near 1, so the R-squared will be close to 0. In that case, the line isn't doing any good as a prediction aid.

The next statistic we'll calculate is s, the standard error. s may be considered an estimate of the standard deviation of the errors from the true line.

Move the cell selector to I13. Type:

`=sqrt(`

sqrt is the abbreviation for the square root function.
It takes the square root of whatever is inside the parentheses.

$$ s = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-2}} $$

Click on I10, or use the arrow keys to move the cell selector to I10, or just type `I10`

I10 has the sum of squared residuals,

$$ \sum_{i=i}^{N}(Y_i - \hat{Y}_i)^2 $$

Then press: /

We want to divide by what is called the "degrees of freedom".

The degrees of freedom in a simple regression is the number of observations (7 here) minus 2. The 2 is because our a simple regression equation has two parameters, the intercept α and the slope β.

Type (after the /):

`(count(e2:e8)-2))`

The count function counts the number of non-blank entries in the range of cells you specify.

We could just type a 5 for the denominator, but it's better to do a formula. Using a formula allows you to change the size of your data set and still get the right answer, without having to remember to change this cell entry.

When finished, the formula for this cell is
`=SQRT(I10/(COUNT(E2:E8)-2))`

Please don't be discouraged if it takes a couple of tries to get this formula right. Nested parentheses can be tricky. There are three left and three right parentheses. One set of parentheses is for the square root, one set is for the denominator of the fraction, and one is for the count function. Excel color-codes parentheses, which helps.

When you have it right, press $\boxed{\text{Enter} \hookleftarrow}$

Completes the formula for s.

Now we put in some statistics that apply specifically to the slope coefficient.

Move the cell selector to D12 and type:
`t coeff=0`

This label emphasizes that the t value here will be for testing the hypothesis that the true slope is 0.

Now move the cell selector to D13.

The t formula will go here.

The t-statistic we'll be using is for a two-tailed test of the hypothesis that the slope of the true line is 0. If the true line has 0 slope, then the true line is horizontal, and there is no linear relationship between X and Y.

Here is the t formula:

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

To put this in D13, let us see how many pieces of this we already have. Beta-hat is in B13. The sum under the square root sign is in E10. *s* is in I13. Good! We have all the pieces we need.

We can rewrite the formula, putting in the cell references:

$$D13 = \frac{B13\sqrt{E10}}{I13}$$

We can type this into cell D13:

`=B13*SQRT(E10)/I13`   and we have our t statistic!

t is our hypothesis tester. We can compare it with the critical t-value that we get from a t-table.   You can find a t-table in the file 716-tables for hypothesis tests.pdf, which you can download using the link on the syllabus web page.

We have seven observations, so our number of degrees of freedom is 5. (N is 7, so N-2 is 5.) According to the t-table, the critical value in the column for 0.05 and the row for 5 is 2.57. That means that a t-value in our spreadsheet of at least 2.57 is needed for significance at the 5% level.   In the table, the critical value in the column for 0.01 and the row for 5 is 4.03. This means that a t-value in our spreadsheet of greater than 4.03 is needed for significance at the 1% level.

The t-value from my spreadsheet is 6.2 (from the picture on page 2). This is bigger than 2.57. Therefore I can reject the hypothesis that the true slope is 0, at the 5% significance level. Also, since 6.2 is bigger than 4.03, I can reject the hypothesis that the true slope is 0 at the 1% significance level. How did your t-value come out? (I'll ask you to formally report your t-value in the next section.)

Let's see what we can learn from the t formula:

1.   The E10 in the formula means that our t-value is higher when our X values are more spread out. The more different the values of X are, the easier it is to see if Y depends on X. Does this make sense intuitively? If you wanted to test the effect of chocolate consumption on acne, you would feed different amounts of chocolate to different people. If you fed everyone the same amount, you would not learn much about the effect of chocolate.

2.   The B13 says that the steeper the slope that you observe, the more likely it is to be statistically significant.

3.   I13 appears as a divisor. The further the observed points are from the line we draw, the <u>less</u> confident we are that there is a true relationship between our variables.

For completeness, we should also put in the standard error of beta-hat, the slope coefficient.

Go to cell C12 and type   to label the cell below
`Std Err`

The formula for the standard error of beta-hat is $\dfrac{s}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$, which is $\dfrac{I13}{\sqrt{E10}}$

In cell C13, type `=i13/sqrt(e10)` `Enter ↵`

Notice that the t number equals the coefficient divided by its standard error.

This spreadsheet is done!

The formulas in the completed spreadsheet are shown here. Ctrl+` shows you all the formulas at once. Alternatively, you can see any individual cell's formula by clicking on the cell and looking at the space just above the column letters. I did that below for cell I13. That formula is too wide to show in the cell, but you can see it on the edit line at the top.

I13		fx	=SQRT(I10/(COUNT(E2:E8)-2))							
	A	B	C	D	E	F	G	H	I	J
1		X	Y	Xdev	XdevSq	Xdev*Y	Pred	Resid	ResidSq	YdevSq
2		100	40	=B2-$B$11	=D2^2	=D2*C2	=$B$14+$B$13*B2	=C2-G2	=H2*H2	=(C2-$C$11)^2
3		200	47	=B3-$B$11	=D3^2	=D3*C3	=$B$14+$B$13*B3	=C3-G3	=H3*H3	=(C3-$C$11)^2
4		300	43	=B4-$B$11	=D4^2	=D4*C4	=$B$14+$B$13*B4	=C4-G4	=H4*H4	=(C4-$C$11)^2
5		400	67	=B5-$B$11	=D5^2	=D5*C5	=$B$14+$B$13*B5	=C5-G5	=H5*H5	=(C5-$C$11)^2
6		500	62	=B6-$B$11	=D6^2	=D6*C6	=$B$14+$B$13*B6	=C6-G6	=H6*H6	=(C6-$C$11)^2
7		600	72	=B7-$B$11	=D7^2	=D7*C7	=$B$14+$B$13*B7	=C7-G7	=H7*H7	=(C7-$C$11)^2
8		700	82	=B8-$B$11	=D8^2	=D8*C8	=$B$14+$B$13*B8	=C8-G8	=H8*H8	=(C8-$C$11)^2
9										
10	Sum	=SUM(B2:B8)	=SUM(C2:C8)	=SUM(D2:D8)	=SUM(E2:E8)	=SUM(F2:F8)	=SUM(G2:G8)	=SUM(H2:H8)	=SUM(I2:I8)	=SUM(J2:J8)
11	Average	=AVERAGE(B2:B8)	=AVERAGE(C2:C8)	=AVERAGE(D2:D8)	=AVERAGE(E2:E8)	=AVERAGE(F2:F8)	=AVERAGE(G2:G8)	=AVERAGE(H2:H8)	=AVERAGE(I2:I8)	=AVERAGE(J2:J8)
12		Coefficients	Std Err	t coeff=0				Regression statistics		
13	Slope	=F10/E10	=I13/SQRT(E10)	=B13*SQRT(E10)/I13				s	=SQRT(I10/(COUNT(	
14	Intercept	=C11-B13*B11						R-squared	=1-I10/J10	

Saving your work

Save your work by clicking the round button and then picking Save As. (Other versions: Select <u>F</u>ile, then Save <u>A</u>s.) Using Save As lets you keep your Assignments 1 and 1A separate.

Now you have a template that can do regressions and give you test statistics, just like SAS! Wow!

Interpretation

Based on the t-value you got, in comparison with the critical t-value of 2.57 at the 5% significance level, what can you say about whether or not the true slop coefficient β is equal to 0?

Write a paragraph in which you describe the assumptions you have to make about the error e in the equation $Y = \alpha + \beta X + e$ in order to make your statement.

Maybe an example of Type I error in hypothesis testing – one more part to assignment 1A.

Once your spreadsheet is saved, there is one more thing I would like you to do with it. This will let you see how the template works with new data. It also will illustrate the possibility of Type I error in hypothesis testing.

Minimize Excel by left-clicking on the ⬁ near the window's upper right corner. You can now see the desktop again.

Start your web browser and go to http://sambaker.com/courses/J716/Data1A.html
Copy, print, or write down the new data.

Click on the Excel button on the task bar at the bottom of your screen, to get your spreadsheet back on the

screen.

If you copied the new data to the clipboard, paste the new data on top of the old data in columns B and C.

If you printed or wrote down the new Y values, type them into cells C2 through C8. Type right over the old values.

When all seven Y values are replaced, write down the new slope, intercept, R^2, and t-value. Based on your t-value, is your slope significantly different from 0?

Save this spreadsheet under a different name.

Assignment 1A checklist:

1.   Turn in a copy of your new spreadsheet with the original data.

2.   Report the slope, intercept, R^2, and t-value you got with the original data. What do you conclude about the true β for these data? Include your supporting paragraph (described above, under Interpretation).

3.   Turn in a copy of your new spreadsheet with the Last Part data.

4.   Report the slope, intercept, R^2, and t-value you got with the Last part data. What do you conclude about the true β for these data?

The syllabus has a link to 716-Philosophy of Hypothesis Testing.pdf. This delves into the philosophy of hypothesis testing, which may help you with your supporting paragraph in item 2.